aws

Amazon Sagemaker API Reference

Amazon SageMaker

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon SageMaker: Amazon Sagemaker API Reference

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Table of Contents

Welcome 1
Amazon SageMaker Service 1
Amazon SageMaker Runtime 1
Amazon Sagemaker Edge Manager 1
Amazon SageMaker Feature Store Runtime 1
Amazon SageMaker geospatial capabilities 1
Amazon SageMaker Metrics Service 2
Amazon SageMaker Job Runtime Service 2
Actions 3
Amazon SageMaker Service 18
AddAssociation 32
AddTags 36
AssociateTrialComponent 39
AttachClusterNodeVolume 42
BatchAddClusterNodes 46
BatchDeleteClusterNodes 50
BatchDescribeModelPackage 55
BatchRebootClusterNodes 60
BatchReplaceClusterNodes 65
CreateAction 71
CreateAIBenchmarkJob 76
CreateAIRecommendationJob 81
CreateAIWorkloadConfig 87
CreateAlgorithm 91
CreateApp 101
CreateAppImageConfig 106
CreateArtifact 110
CreateAutoMLJob 114
CreateAutoMLJobV2 121
CreateCluster 128
CreateClusterSchedulerConfig 138
CreateCodeRepository 142
CreateCompilationJob 145
CreateComputeQuota 151

CreateContext ... 156
CreateDataQualityJobDefinition ... 160
CreateDeviceFleet ... 166
CreateDomain ... 170
CreateEdgeDeploymentPlan ... 185
CreateEdgeDeploymentStage ... 189
CreateEdgePackagingJob ... 192
CreateEndpoint ... 196
CreateEndpointConfig ... 202
CreateExperiment ... 213
CreateFeatureGroup ... 217
CreateFlowDefinition ... 225
CreateHub ... 230
CreateHubContentPresignedUrls ... 234
CreateHubContentReference ... 238
CreateHumanTaskUi ... 242
CreateHyperParameterTuningJob ... 245
CreateImage ... 259
CreateImageVersion ... 263
CreateInferenceComponent ... 269
CreateInferenceExperiment ... 275
CreateInferenceRecommendationsJob ... 282
CreateJob ... 288
CreateLabelingJob ... 293
CreateMlflowApp ... 302
CreateMlflowTrackingServer ... 307
CreateModel ... 312
CreateModelBiasJobDefinition ... 320
CreateModelCard ... 326
CreateModelCardExportJob ... 330
CreateModelExplainabilityJobDefinition ... 334
CreateModelPackage ... 340
CreateModelPackageGroup ... 357
CreateModelQualityJobDefinition ... 360
CreateMonitoringSchedule ... 366
CreateNotebookInstance ... 371

CreateNotebookInstanceLifecycleConfig 381
CreateOptimizationJob 385
CreatePartnerApp 392
CreatePartnerAppPresignedUrl 398
CreatePipeline 401
CreatePresignedDomainUrl 406
CreatePresignedMlflowAppUrl 411
CreatePresignedMlflowTrackingServerUrl 414
CreatePresignedNotebookInstanceUrl 417
CreateProcessingJob 420
CreateProject 428
CreateSpace 432
CreateStudioLifecycleConfig 438
CreateTrainingJob 441
CreateTrainingPlan 458
CreateTransformJob 462
CreateTrial 470
CreateTrialComponent 474
CreateUserProfile 480
CreateWorkforce 489
CreateWorkteam 494
DeleteAction 499
DeleteAIBenchmarkJob 501
DeleteAIRecommendationJob 503
DeleteAIWorkloadConfig 505
DeleteAlgorithm 508
DeleteApp 510
DeleteAppImageConfig 513
DeleteArtifact 515
DeleteAssociation 518
DeleteCluster 521
DeleteClusterSchedulerConfig 524
DeleteCodeRepository 526
DeleteCompilationJob 528
DeleteComputeQuota 530
DeleteContext 532

DeleteDataQualityJobDefinition 534
DeleteDeviceFleet 536
DeleteDomain 538
DeleteEdgeDeploymentPlan 540
DeleteEdgeDeploymentStage 542
DeleteEndpoint 544
DeleteEndpointConfig 546
DeleteExperiment 548
DeleteFeatureGroup 550
DeleteFlowDefinition 552
DeleteHub 554
DeleteHubContent 556
DeleteHubContentReference 559
DeleteHumanTaskUi 562
DeleteHyperParameterTuningJob 564
DeleteImage 566
DeleteImageVersion 568
DeleteInferenceComponent 571
DeleteInferenceExperiment 573
DeleteJob 576
DeleteMlflowApp 579
DeleteMlflowTrackingServer 581
DeleteModel 584
DeleteModelBiasJobDefinition 586
DeleteModelCard 588
DeleteModelExplainabilityJobDefinition 590
DeleteModelPackage 592
DeleteModelPackageGroup 594
DeleteModelPackageGroupPolicy 596
DeleteModelQualityJobDefinition 598
DeleteMonitoringSchedule 600
DeleteNotebookInstance 602
DeleteNotebookInstanceLifecycleConfig 604
DeleteOptimizationJob 606
DeletePartnerApp 608
DeletePipeline 611

DeleteProcessingJob 614
DeleteProject 616
DeleteSpace 618
DeleteStudioLifecycleConfig 620
DeleteTags 622
DeleteTrainingJob 625
DeleteTrial 627
DeleteTrialComponent 629
DeleteUserProfile 632
DeleteWorkforce 634
DeleteWorkteam 636
DeregisterDevices 638
DescribeAction 640
DescribeAIBenchmarkJob 646
DescribeAIRecommendationJob 651
DescribeAIWorkloadConfig 658
DescribeAlgorithm 662
DescribeApp 672
DescribeAppImageConfig 679
DescribeArtifact 683
DescribeAutoMLJob 688
DescribeAutoMLJobV2 697
DescribeCluster 706
DescribeClusterEvent 715
DescribeClusterNode 718
DescribeClusterSchedulerConfig 723
DescribeCodeRepository 729
DescribeCompilationJob 732
DescribeComputeQuota 739
DescribeContext 745
DescribeDataQualityJobDefinition 750
DescribeDevice 756
DescribeDeviceFleet 761
DescribeDomain 765
DescribeEdgeDeploymentPlan 780
DescribeEdgePackagingJob 785

DescribeEndpoint ... 790
DescribeEndpointConfig ... 803
DescribeExperiment ... 811
DescribeFeatureGroup ... 815
DescribeFeatureMetadata ... 823
DescribeFlowDefinition ... 827
DescribeHub ... 832
DescribeHubContent ... 836
DescribeHumanTaskUi ... 843
DescribeHyperParameterTuningJob ... 846
DescribeImage ... 861
DescribeImageVersion ... 865
DescribeInferenceComponent ... 871
DescribeInferenceExperiment ... 878
DescribeInferenceRecommendationsJob ... 885
DescribeJob ... 893
DescribeJobSchemaVersion ... 899
DescribeLabelingJob ... 902
DescribeLineageGroup ... 910
DescribeMlflowApp ... 914
DescribeMlflowTrackingServer ... 920
DescribeModel ... 926
DescribeModelBiasJobDefinition ... 933
DescribeModelCard ... 939
DescribeModelCardExportJob ... 945
DescribeModelExplainabilityJobDefinition ... 949
DescribeModelPackage ... 955
DescribeModelPackageGroup ... 973
DescribeModelQualityJobDefinition ... 977
DescribeMonitoringSchedule ... 983
DescribeNotebookInstance ... 990
DescribeNotebookInstanceLifecycleConfig ... 999
DescribeOptimizationJob ... 1002
DescribePartnerApp ... 1009
DescribePipeline ... 1016
DescribePipelineDefinitionForExecution ... 1022

DescribePipelineExecution 1025
DescribeProcessingJob 1031
DescribeProject 1039
DescribeReservedCapacity 1044
DescribeSpace 1049
DescribeStudioLifecycleConfig 1056
DescribeSubscribedWorkteam 1059
DescribeTrainingJob 1061
DescribeTrainingPlan 1080
DescribeTrainingPlanExtensionHistory 1086
DescribeTransformJob 1090
DescribeTrial 1098
DescribeTrialComponent 1103
DescribeUserProfile 1110
DescribeWorkforce 1119
DescribeWorkteam 1122
DetachClusterNodeVolume 1125
DisableSagemakerServicecatalogPortfolio 1129
DisassociateTrialComponent 1130
EnableSagemakerServicecatalogPortfolio 1133
ExtendTrainingPlan 1134
GetDeviceFleetReport 1137
GetLineageGroupPolicy 1141
GetModelPackageGroupPolicy 1144
GetSagemakerServicecatalogPortfolioStatus 1146
GetScalingConfigurationRecommendation 1148
GetSearchSuggestions 1153
ImportHubContent 1156
ListActions 1162
ListAIBenchmarkJobs 1166
ListAIRecommendationJobs 1170
ListAIWorkloadConfigs 1174
ListAlgorithms 1178
ListAliases 1182
ListAppImageConfigs 1186
ListApps 1191

ListArtifacts 1196
ListAssociations 1201
ListAutoMLJobs 1206
ListCandidatesForAutoMLJob 1211
ListClusterEvents 1216
ListClusterNodes 1221
ListClusters 1226
ListClusterSchedulerConfigs 1231
ListCodeRepositories 1236
ListCompilationJobs 1241
ListComputeQuotas 1246
ListContexts 1252
ListDataQualityJobDefinitions 1256
ListDeviceFleets 1260
ListDevices 1264
ListDomains 1268
ListEdgeDeploymentPlans 1271
ListEdgePackagingJobs 1276
ListEndpointConfigs 1281
ListEndpoints 1285
ListExperiments 1290
ListFeatureGroups 1294
ListFlowDefinitions 1298
ListHubContents 1302
ListHubContentVersions 1307
ListHubs 1312
ListHumanTaskUis 1316
ListHyperParameterTuningJobs 1319
ListImages 1324
ListImageVersions 1328
ListInferenceComponents 1333
ListInferenceExperiments 1338
ListInferenceRecommendationsJobs 1343
ListInferenceRecommendationsJobSteps 1348
ListJobs 1353
ListJobSchemaVersions 1358

ListLabelingJobs 1361
ListLabelingJobsForWorkteam 1366
ListLineageGroups 1371
ListMlflowApps 1375
ListMlflowTrackingServers 1380
ListModelBiasJobDefinitions 1385
ListModelCardExportJobs 1389
ListModelCards 1394
ListModelCardVersions 1398
ListModelExplainabilityJobDefinitions 1402
ListModelMetadata 1406
ListModelPackageGroups 1409
ListModelPackages 1413
ListModelQualityJobDefinitions 1418
ListModels 1422
ListMonitoringAlertHistory 1426
ListMonitoringAlerts 1431
ListMonitoringExecutions 1435
ListMonitoringSchedules 1441
ListNotebookInstanceLifecycleConfigs 1446
ListNotebookInstances 1451
ListOptimizationJobs 1457
ListPartnerApps 1462
ListPipelineExecutions 1465
ListPipelineExecutionSteps 1469
ListPipelineParametersForExecution 1476
ListPipelines 1479
ListPipelineVersions 1483
ListProcessingJobs 1487
ListProjects 1492
ListResourceCatalogs 1496
ListSpaces 1500
ListStageDevices 1504
ListStudioLifecycleConfigs 1508
ListSubscribedWorkteams 1513
ListTags 1516

ListTrainingJobs 1519
ListTrainingJobsForHyperParameterTuningJob 1525
ListTrainingPlans 1529
ListTransformJobs 1534
ListTrialComponents 1539
ListTrials 1545
ListUltraServersByReservedCapacity 1550
ListUserProfiles 1553
ListWorkforces 1557
ListWorkteams 1561
PutModelPackageGroupPolicy 1565
QueryLineage 1568
RegisterDevices 1573
RenderUiTemplate 1576
RetryPipelineExecution 1580
Search 1583
SearchTrainingPlanOfferings 1651
SendPipelineExecutionStepFailure 1657
SendPipelineExecutionStepSuccess 1660
StartClusterHealthCheck 1663
StartEdgeDeploymentStage 1666
StartInferenceExperiment 1668
StartMlflowTrackingServer 1671
StartMonitoringSchedule 1674
StartNotebookInstance 1676
StartPipelineExecution 1678
StartSession 1683
StopAIBenchmarkJob 1686
StopAIRecommendationJob 1688
StopAutoMLJob 1690
StopCompilationJob 1692
StopEdgeDeploymentStage 1694
StopEdgePackagingJob 1696
StopHyperParameterTuningJob 1698
StopInferenceExperiment 1700
StopInferenceRecommendationsJob 1704

StopJob 1706
StopLabelingJob 1708
StopMlflowTrackingServer 1710
StopMonitoringSchedule 1713
StopNotebookInstance 1715
StopOptimizationJob 1717
StopPipelineExecution 1719
StopProcessingJob 1722
StopTrainingJob 1724
StopTransformJob 1726
UpdateAction 1728
UpdateAppImageConfig 1732
UpdateArtifact 1736
UpdateCluster 1740
UpdateClusterSchedulerConfig 1749
UpdateClusterSoftware 1753
UpdateCodeRepository 1758
UpdateComputeQuota 1761
UpdateContext 1766
UpdateDeviceFleet 1770
UpdateDevices 1773
UpdateDomain 1775
UpdateEndpoint 1788
UpdateEndpointWeightsAndCapacities 1793
UpdateExperiment 1796
UpdateFeatureGroup 1799
UpdateFeatureMetadata 1803
UpdateHub 1806
UpdateHubContent 1809
UpdateHubContentReference 1815
UpdateImage 1819
UpdateImageVersion 1823
UpdateInferenceComponent 1829
UpdateInferenceComponentRuntimeConfig 1835
UpdateInferenceExperiment 1838
UpdateMlflowApp 1843

UpdateMlflowTrackingServer 1847
UpdateModelCard 1851
UpdateModelPackage 1855
UpdateMonitoringAlert 1864
UpdateMonitoringSchedule 1868
UpdateNotebookInstance 1873
UpdateNotebookInstanceLifecycleConfig 1881
UpdatePartnerApp 1884
UpdatePipeline 1889
UpdatePipelineExecution 1893
UpdatePipelineVersion 1896
UpdateProject 1900
UpdateSpace 1904
UpdateTrainingJob 1909
UpdateTrial 1913
UpdateTrialComponent 1916
UpdateUserProfile 1922
UpdateWorkforce 1930
UpdateWorkteam 1936
Amazon SageMaker Runtime 1940
InvokeEndpoint 1941
InvokeEndpointAsync 1950
InvokeEndpointWithResponseStream 1956
Amazon Sagemaker Edge Manager 1964
GetDeployments 1965
GetDeviceRegistration 1968
SendHeartbeat 1971
Amazon SageMaker Feature Store Runtime 1974
BatchGetRecord 1976
DeleteRecord 1980
GetRecord 1984
PutRecord 1988
Amazon SageMaker geospatial capabilities 1991
DeleteEarthObservationJob 1993
DeleteVectorEnrichmentJob 1996
ExportEarthObservationJob 1999

ExportVectorEnrichmentJob ... 2005
GetEarthObservationJob ... 2010
GetRasterDataCollection ... 2016
GetTile ... 2020
GetVectorEnrichmentJob ... 2024
ListEarthObservationJobs ... 2030
ListRasterDataCollections ... 2035
ListTagsForResource ... 2039
ListVectorEnrichmentJobs ... 2042
SearchRasterDataCollection ... 2047
StartEarthObservationJob ... 2052
StartVectorEnrichmentJob ... 2059
StopEarthObservationJob ... 2066
StopVectorEnrichmentJob ... 2069
TagResource ... 2072
UntagResource ... 2075
Amazon SageMaker Metrics Service ... 2077
BatchGetMetrics ... 2078
BatchPutMetrics ... 2081
Amazon SageMaker Job Runtime Service ... 2083
CompleteRollout ... 2084
Sample ... 2088
SampleWithResponseStream ... 2092
UpdateReward ... 2096
Data Types ... 2100
Amazon SageMaker Service ... 2129
AcceleratorPartitionConfig ... 2156
ActionSource ... 2158
ActionSummary ... 2160
AddClusterNodeSpecification ... 2162
AdditionalEnis ... 2165
AdditionalInferenceSpecificationDefinition ... 2166
AdditionalModelDataSource ... 2171
AdditionalS3DataSource ... 2172
AgentVersion ... 2174
AIBenchmarkEndpoint ... 2175

AIBenchmarkInferenceComponent 2177
AIBenchmarkJobSummary 2178
AIBenchmarkNetworkConfig 2180
AIBenchmarkOutputConfig 2181
AIBenchmarkOutputResult 2182
AIBenchmarkTarget 2184
AICapacityReservationConfig 2185
AICloudWatchLogs 2186
AIDatasetConfig 2187
AIMlflowConfig 2188
AIModelSource 2190
AIModelSourceS3 2191
AIRecommendation 2192
AIRecommendationComputeSpec 2194
AIRecommendationConstraint 2196
AIRecommendationDeploymentConfiguration 2197
AIRecommendationDeploymentS3Channel 2199
AIRecommendationInferenceSpecification 2200
AIRecommendationInstanceDetail 2201
AIRecommendationJobSummary 2203
AIRecommendationModelDetails 2205
AIRecommendationOptimizationDetail 2207
AIRecommendationOutputConfig 2208
AIRecommendationOutputResult 2210
AIRecommendationPerformanceMetric 2212
AIRecommendationPerformanceTarget 2214
AIWorkloadConfigs 2215
AIWorkloadConfigSummary 2216
AIWorkloadDataSource 2218
AIWorkloadInputDataConfig 2219
AIWorkloadS3DataSource 2220
Alarm 2221
AlarmDetails 2222
AlgorithmSpecification 2223
AlgorithmStatusDetails 2228
AlgorithmStatusItem 2229

AlgorithmSummary 2231
AlgorithmValidationProfile 2233
AlgorithmValidationSpecification 2235
AmazonQSettings 2236
AnnotationConsolidationConfig 2237
AppDetails 2256
AppImageConfigDetails 2259
AppLifecycleManagement 2261
AppSpecification 2262
ArtifactSource 2264
ArtifactSourceType 2265
ArtifactSummary 2266
AssociationInfo 2268
AssociationSummary 2269
AsyncInferenceClientConfig 2272
AsyncInferenceConfig 2273
AsyncInferenceNotificationConfig 2274
AsyncInferenceOutputConfig 2276
AthenaDatasetDefinition 2278
AuthorizedUrl 2281
AutoMLAlgorithmConfig 2283
AutoMLCandidate 2285
AutoMLCandidateGenerationConfig 2288
AutoMLCandidateStep 2291
AutoMLChannel 2293
AutoMLComputeConfig 2296
AutoMLContainerDefinition 2298
AutoMLDataSource 2300
AutoMLDataSplitConfig 2301
AutoMLJobArtifacts 2302
AutoMLJobChannel 2303
AutoMLJobCompletionCriteria 2305
AutoMLJobConfig 2307
AutoMLJobObjective 2309
AutoMLJobStepMetadata 2311
AutoMLJobSummary 2312

AutoMLOutputDataConfig 2315
AutoMLPartialFailureReason 2316
AutoMLProblemTypeConfig 2317
AutoMLProblemTypeResolvedAttributes 2319
AutoMLResolvedAttributes 2320
AutoMLS3DataSource 2321
AutoMLSecurityConfig 2323
AutoParameter 2325
AutoRollbackConfig 2327
Autotune 2328
AvailableUpgrade 2330
BaseModel 2332
BatchAddClusterNodesError 2334
BatchDataCaptureConfig 2337
BatchDeleteClusterNodeLogicalIdsError 2339
BatchDeleteClusterNodesError 2341
BatchDescribeModelPackageError 2343
BatchDescribeModelPackageSummary 2344
BatchRebootClusterNodeLogicalIdsError 2347
BatchRebootClusterNodesError 2349
BatchReplaceClusterNodeLogicalIdsError 2351
BatchReplaceClusterNodesError 2353
BatchTransformInput 2355
BedrockCustomModelDeploymentMetadata 2359
BedrockCustomModelMetadata 2360
BedrockModelImportMetadata 2361
BedrockProvisionedModelThroughputMetadata 2362
BestObjectiveNotImproving 2363
Bias 2364
BlueGreenUpdatePolicy 2365
CacheHitResult 2367
CallbackStepMetadata 2368
CandidateArtifactLocations 2370
CandidateGenerationConfig 2372
CandidateProperties 2374
CanvasAppSettings 2375

CapacityReservation 2377
CapacitySize 2378
CapacitySizeConfig 2380
CaptureContentTypeHeader 2381
CaptureOption 2383
CategoricalParameter 2384
CategoricalParameterRange 2385
CategoricalParameterRangeSpecification 2387
CfnCreateTemplateProvider 2388
CfnStackCreateParameter 2390
CfnStackDetail 2391
CfnStackParameter 2393
CfnStackUpdateParameter 2394
CfnTemplateProviderDetail 2395
CfnUpdateTemplateProvider 2397
Channel 2399
ChannelSpecification 2402
CheckpointConfig 2405
ClarifyCheckStepMetadata 2407
ClarifyExplainerConfig 2410
ClarifyInferenceConfig 2412
ClarifyShapBaselineConfig 2417
ClarifyShapConfig 2419
ClarifyTextConfig 2421
ClusterAutoScalingConfig 2423
ClusterAutoScalingConfigOutput 2424
ClusterCapacityRequirements 2426
ClusterEbsVolumeConfig 2427
ClusterEventDetail 2429
ClusterEventSummary 2432
ClusterFsxLustreConfig 2435
ClusterFsxOpenZfsConfig 2437
ClusterInstanceGroupDetails 2438
ClusterInstanceGroupSpecification 2447
ClusterInstancePlacement 2454
ClusterInstanceRequirementDetails 2455

ClusterInstanceRequirements 2458
ClusterInstanceStatusDetails 2460
ClusterInstanceStorageConfig 2461
ClusterInstanceTypeDetail 2463
ClusterKubernetesConfig 2465
ClusterKubernetesConfigDetails 2467
ClusterKubernetesConfigNodeDetails 2469
ClusterKubernetesTaint 2471
ClusterLifeCycleConfig 2473
ClusterMetadata 2475
ClusterNetworkInterface 2477
ClusterNetworkInterfaceDetails 2478
ClusterNodeDetails 2479
ClusterNodeSummary 2485
ClusterOnDemandOptions 2489
ClusterOrchestrator 2490
ClusterOrchestratorEksConfig 2491
ClusterOrchestratorSlurmConfig 2492
ClusterRestrictedInstanceGroupDetails 2493
ClusterRestrictedInstanceGroupsConfig 2498
ClusterRestrictedInstanceGroupsConfigOutput 2499
ClusterRestrictedInstanceGroupSpecification 2500
ClusterSchedulerConfigSummary 2504
ClusterSharedEnvironmentConfig 2507
ClusterSharedEnvironmentConfigDetails 2508
ClusterSlurmConfig 2510
ClusterSlurmConfigDetails 2511
ClusterSpotOptions 2512
ClusterSummary 2513
ClusterTieredStorageConfig 2515
CodeEditorAppImageConfig 2517
CodeEditorAppSettings 2518
CodeRepository 2520
CodeRepositorySummary 2521
CognitoConfig 2523
CognitoMemberDefinition 2525

CollectionConfig 2527
CollectionConfiguration 2528
CompilationJobSummary 2530
ComputeQuotaConfig 2533
ComputeQuotaResourceConfig 2535
ComputeQuotaSummary 2538
ComputeQuotaTarget 2541
ConditionStepMetadata 2542
ContainerConfig 2543
ContainerDefinition 2545
ContainerMetricsConfig 2551
ContextSource 2552
ContextSummary 2554
ContinuousParameterRange 2556
ContinuousParameterRangeSpecification 2559
ConvergenceDetected 2560
CreateTemplateProvider 2561
CustomFileSystem 2562
CustomFileSystemConfig 2564
CustomImage 2566
CustomizedMetricSpecification 2568
CustomPosixUserConfig 2569
DataCaptureConfig 2570
DataCaptureConfigSummary 2573
DataCatalogConfig 2575
DataProcessing 2577
DataQualityAppSpecification 2579
DataQualityBaselineConfig 2582
DataQualityJobInput 2584
DatasetDefinition 2585
DatasetSource 2587
DataSource 2588
DebugHookConfig 2589
DebugRuleConfiguration 2591
DebugRuleEvaluationStatus 2595
DefaultEbsStorageSettings 2597

DefaultSpaceSettings 2598
DefaultSpaceStorageSettings 2601
DeployedImage 2602
DeploymentConfig 2604
DeploymentConfiguration 2606
DeploymentRecommendation 2608
DeploymentStage 2610
DeploymentStageStatusSummary 2612
DerivedInformation 2614
DesiredWeightAndCapacity 2615
Device 2617
DeviceDeploymentSummary 2619
DeviceFleetSummary 2623
DeviceSelectionConfig 2625
DeviceStats 2627
DeviceSummary 2628
DirectDeploySettings 2631
DockerSettings 2632
DomainDetails 2634
DomainSettings 2637
DomainSettingsForUpdate 2640
DriftCheckBaselines 2643
DriftCheckBias 2645
DriftCheckExplainability 2646
DriftCheckModelDataQuality 2647
DriftCheckModelQuality 2648
DynamicScalingConfiguration 2649
EbsStorageSettings 2651
Ec2CapacityReservation 2652
Edge 2654
EdgeDeploymentConfig 2656
EdgeDeploymentModelConfig 2657
EdgeDeploymentPlanSummary 2658
EdgeDeploymentStatus 2661
EdgeModel 2663
EdgeModelStat 2665

EdgeModelSummary 2667
EdgeOutputConfig 2668
EdgePackagingJobSummary 2670
EdgePresetDeploymentOutput 2673
EFSFileSystem 2675
EFSFileSystemConfig 2676
EmrServerlessComputeConfig 2678
EmrServerlessSettings 2679
EmrSettings 2680
EMRStepMetadata 2682
Endpoint 2684
EndpointConfigStepMetadata 2687
EndpointConfigSummary 2688
EndpointInfo 2690
EndpointInput 2691
EndpointInputConfiguration 2694
EndpointMetadata 2698
EndpointOutputConfiguration 2700
EndpointPerformance 2704
EndpointStepMetadata 2705
EndpointSummary 2706
EnvironmentConfig 2709
EnvironmentConfigDetails 2710
EnvironmentParameter 2711
EnvironmentParameterRanges 2712
ErrorInfo 2713
EventDetails 2714
EventMetadata 2715
Experiment 2717
ExperimentConfig 2720
ExperimentSource 2722
ExperimentSummary 2723
Explainability 2725
ExplainerConfig 2726
FailStepMetadata 2727
FeatureDefinition 2728

FeatureGroup 2730
FeatureGroupSummary 2735
FeatureMetadata 2737
FeatureParameter 2740
FileSource 2741
FileSystemConfig 2743
FileSystemDataSource 2745
Filter 2747
FinalAutoMLJobObjectiveMetric 2751
FinalHyperParameterTuningJobObjectiveMetric 2753
FlowDefinitionOutputConfig 2755
FlowDefinitionSummary 2757
FSxLustreConfig 2759
FSxLustreFileSystem 2760
FSxLustreFileSystemConfig 2761
GenerativeAiSettings 2763
GitConfig 2764
GitConfigForUpdate 2766
HiddenSageMakerImage 2767
HolidayConfigAttributes 2768
HubAccessConfig 2769
HubContentDependency 2770
HubContentInfo 2771
HubInfo 2775
HubS3StorageConfig 2778
HumanLoopActivationConditionsConfig 2779
HumanLoopActivationConfig 2780
HumanLoopConfig 2781
HumanLoopRequestSource 2788
HumanTaskConfig 2789
HumanTaskUiSummary 2811
HyperbandStrategyConfig 2813
HyperParameterAlgorithmSpecification 2815
HyperParameterSpecification 2818
HyperParameterTrainingJobDefinition 2820
HyperParameterTrainingJobSummary 2826

HyperParameterTuningInstanceConfig 2830
HyperParameterTuningJobCompletionDetails 2833
HyperParameterTuningJobConfig 2834
HyperParameterTuningJobConsumedResources 2837
HyperParameterTuningJobObjective 2838
HyperParameterTuningJobSearchEntity 2840
HyperParameterTuningJobStrategyConfig 2845
HyperParameterTuningJobSummary 2846
HyperParameterTuningJobWarmStartConfig 2849
HyperParameterTuningResourceConfig 2851
IamIdentity 2856
IamPolicyConstraints 2857
IdentityProviderOAuthSetting 2859
IdleSettings 2861
Image 2863
ImageClassificationJobConfig 2866
ImageConfig 2867
ImageVersion 2868
InferenceComponentAvailabilityZoneBalance 2870
InferenceComponentCapacitySize 2872
InferenceComponentComputeResourceRequirements 2874
InferenceComponentContainerSpecification 2876
InferenceComponentContainerSpecificationSummary 2878
InferenceComponentDataCacheConfig 2880
InferenceComponentDataCacheConfigSummary 2881
InferenceComponentDeploymentConfig 2882
InferenceComponentMetadata 2883
InferenceComponentPlacementStatus 2884
InferenceComponentRollingUpdatePolicy 2887
InferenceComponentRuntimeConfig 2889
InferenceComponentRuntimeConfigSummary 2890
InferenceComponentSchedulingConfig 2892
InferenceComponentSpecification 2894
InferenceComponentSpecificationSummary 2899
InferenceComponentStartupParameters 2904
InferenceComponentSummary 2905

InferenceExecutionConfig 2908
InferenceExperimentDataStorageConfig 2909
InferenceExperimentSchedule 2911
InferenceExperimentSummary 2912
InferenceHubAccessConfig 2915
InferenceMetrics 2916
InferenceRecommendation 2917
InferenceRecommendationsJob 2919
InferenceRecommendationsJobStep 2923
InferenceSpecification 2925
InfraCheckConfig 2930
InputConfig 2931
InstanceGroup 2936
InstanceGroupHealthCheckConfiguration 2939
InstanceGroupMetadata 2941
InstanceGroupScalingMetadata 2943
InstanceMetadata 2945
InstanceMetadataServiceConfiguration 2947
InstancePlacementConfig 2948
InstancePool 2949
InstancePoolSummary 2953
InstanceRequirementsEniConfiguration 2956
IntegerParameterRange 2957
IntegerParameterRangeSpecification 2959
Job 2960
JobConfigSchemaVersionSummary 2964
JobSecondaryStatusTransition 2965
JobStepMetadata 2967
JobSummary 2968
JupyterLabAppImageConfig 2971
JupyterLabAppSettings 2972
JupyterServerAppSettings 2975
KendraSettings 2977
KernelGatewayAppSettings 2978
KernelGatewayImageConfig 2980
KernelSpec 2981

LabelCounters 2982
LabelCountersForWorkteam 2984
LabelingJobAlgorithmsConfig 2986
LabelingJobDataAttributes 2988
LabelingJobDataSource 2989
LabelingJobForWorkteamSummary 2991
LabelingJobInputConfig 2993
LabelingJobOutput 2994
LabelingJobOutputConfig 2995
LabelingJobResourceConfig 2997
LabelingJobS3DataSource 2999
LabelingJobSnsDataSource 3000
LabelingJobStoppingConditions 3001
LabelingJobSummary 3003
LambdaStepMetadata 3007
LastUpdateStatus 3008
LineageGroupSummary 3009
LineageMetadata 3011
ManagedConfiguration 3013
MemberDefinition 3014
MetadataProperties 3015
MetricData 3017
MetricDatum 3019
MetricDefinition 3021
MetricsConfig 3023
MetricsEndpoint 3025
MetricSpecification 3027
MetricsSource 3028
MlflowAppSummary 3030
MlflowConfig 3032
MLflowConfiguration 3034
MlflowDetails 3035
Model 3036
ModelAccessConfig 3039
ModelArtifacts 3040
ModelBiasAppSpecification 3041

ModelBiasBaselineConfig 3043
ModelBiasJobInput 3044
ModelCard 3045
ModelCardExportArtifacts 3049
ModelCardExportJobSummary 3050
ModelCardExportOutputConfig 3052
ModelCardSecurityConfig 3053
ModelCardSummary 3054
ModelCardVersionSummary 3056
ModelClientConfig 3058
ModelCompilationConfig 3059
ModelConfiguration 3061
ModelDashboardEndpoint 3063
ModelDashboardIndicatorAction 3065
ModelDashboardModel 3066
ModelDashboardModelCard 3068
ModelDashboardMonitoringSchedule 3071
ModelDataQuality 3074
ModelDataSource 3075
ModelDeployConfig 3076
ModelDeployResult 3078
ModelDigests 3079
ModelExplainabilityAppSpecification 3080
ModelExplainabilityBaselineConfig 3082
ModelExplainabilityJobInput 3083
ModelInfrastructureConfig 3084
ModelInput 3085
ModelLatencyThreshold 3086
ModelLifeCycle 3087
ModelMetadataFilter 3089
ModelMetadataSearchExpression 3090
ModelMetadataSummary 3091
ModelMetrics 3093
ModelPackage 3095
ModelPackageConfig 3103
ModelPackageContainerDefinition 3105

ModelPackageGroup 3110
ModelPackageGroupSummary 3113
ModelPackageModelCard 3115
ModelPackageSecurityConfig 3117
ModelPackageStatusDetails 3118
ModelPackageStatusItem 3119
ModelPackageSummary 3121
ModelPackageValidationProfile 3124
ModelPackageValidationSpecification 3125
ModelQuality 3126
ModelQualityAppSpecification 3127
ModelQualityBaselineConfig 3130
ModelQualityJobInput 3131
ModelQuantizationConfig 3132
ModelRegisterSettings 3134
ModelShardingConfig 3135
ModelSpeculativeDecodingConfig 3137
ModelSpeculativeDecodingTrainingDataSource 3138
ModelStepMetadata 3139
ModelSummary 3140
ModelVariantConfig 3142
ModelVariantConfigSummary 3144
MonitoringAlertActions 3146
MonitoringAlertHistorySummary 3147
MonitoringAlertSummary 3149
MonitoringAppSpecification 3151
MonitoringBaselineConfig 3153
MonitoringClusterConfig 3155
MonitoringConstraintsResource 3158
MonitoringCsvDatasetFormat 3159
MonitoringDatasetFormat 3160
MonitoringExecutionSummary 3161
MonitoringGroundTruthS3Input 3164
MonitoringInput 3165
MonitoringJobDefinition 3166
MonitoringJobDefinitionSummary 3169

MonitoringJsonDatasetFormat 3171
MonitoringNetworkConfig 3172
MonitoringOutput 3174
MonitoringOutputConfig 3175
MonitoringParquetDatasetFormat 3176
MonitoringResources 3177
MonitoringS3Output 3178
MonitoringSchedule 3180
MonitoringScheduleConfig 3183
MonitoringScheduleSummary 3185
MonitoringStatisticsResource 3188
MonitoringStoppingCondition 3189
MultiModelConfig 3190
NeoVpcConfig 3191
NestedFilters 3193
NetworkConfig 3195
NodeAdditionResult 3197
NotebookInstanceLifecycleConfigSummary 3200
NotebookInstanceLifecycleHook 3202
NotebookInstanceSummary 3204
NotificationConfiguration 3209
ObjectiveStatusCounters 3210
OfflineStoreConfig 3212
OfflineStoreStatus 3214
OidcConfig 3215
OidcConfigForResponse 3219
OidcMemberDefinition 3222
OnlineStoreConfig 3223
OnlineStoreConfigUpdate 3225
OnlineStoreSecurityConfig 3226
OptimizationConfig 3228
OptimizationJobModelSource 3230
OptimizationJobModelSourceS3 3231
OptimizationJobOutputConfig 3232
OptimizationJobSummary 3234
OptimizationModelAccessConfig 3237

OptimizationOutput 3238
OptimizationSageMakerModel 3239
OptimizationVpcConfig 3240
OutputConfig 3242
OutputDataConfig 3247
OutputParameter 3249
OwnershipSettings 3250
OwnershipSettingsSummary 3251
ParallelismConfiguration 3252
Parameter 3253
ParameterRange 3254
ParameterRanges 3256
Parent 3258
ParentHyperParameterTuningJob 3260
PartnerAppConfig 3261
PartnerAppMaintenanceConfig 3263
PartnerAppSummary 3264
PendingDeploymentSummary 3266
PendingProductionVariantSummary 3268
Phase 3274
Pipeline 3276
PipelineDefinitionS3Location 3280
PipelineExecution 3282
PipelineExecutionStep 3286
PipelineExecutionStepMetadata 3289
PipelineExecutionSummary 3294
PipelineExperimentConfig 3296
PipelineSummary 3297
PipelineVersion 3300
PipelineVersionSummary 3303
PlacementSpecification 3305
PredefinedMetricSpecification 3306
PresignedUrlAccessConfig 3307
PriorityClass 3308
ProcessingClusterConfig 3309
ProcessingFeatureStoreOutput 3312

ProcessingInput ... 3313
ProcessingJob ... 3315
ProcessingJobStepMetadata ... 3321
ProcessingJobSummary ... 3322
ProcessingOutput ... 3325
ProcessingOutputConfig ... 3327
ProcessingResources ... 3328
ProcessingS3Input ... 3329
ProcessingS3Output ... 3332
ProcessingStoppingCondition ... 3334
ProductionVariant ... 3335
ProductionVariantCapacityReservationConfig ... 3343
ProductionVariantCapacityReservationSummary ... 3345
ProductionVariantCoreDumpConfig ... 3347
ProductionVariantManagedInstanceScaling ... 3349
ProductionVariantManagedInstanceScalingScaleInPolicy ... 3351
ProductionVariantRoutingConfig ... 3353
ProductionVariantServerlessConfig ... 3354
ProductionVariantServerlessUpdateConfig ... 3356
ProductionVariantStatus ... 3357
ProductionVariantSummary ... 3359
ProfilerConfig ... 3363
ProfilerConfigForUpdate ... 3365
ProfilerRuleConfiguration ... 3367
ProfilerRuleEvaluationStatus ... 3371
Project ... 3373
ProjectSummary ... 3377
PropertyNameQuery ... 3379
PropertyNameSuggestion ... 3380
ProvisioningParameter ... 3381
PublicWorkforceTaskPrice ... 3382
QualityCheckStepMetadata ... 3387
QueryFilters ... 3390
RealTimeInferenceConfig ... 3393
RealTimeInferenceRecommendation ... 3396
RecommendationJobCompiledOutputConfig ... 3400

RecommendationJobContainerConfig 3401
RecommendationJobInferenceBenchmark 3404
RecommendationJobInputConfig 3406
RecommendationJobOutputConfig 3410
RecommendationJobPayloadConfig 3412
RecommendationJobResourceLimit 3413
RecommendationJobStoppingConditions 3414
RecommendationJobVpcConfig 3416
RecommendationMetrics 3418
RedshiftDatasetDefinition 3420
RegisterModelStepMetadata 3423
RemoteDebugConfig 3424
RemoteDebugConfigForUpdate 3425
RenderableTask 3426
RenderingError 3427
RepositoryAuthConfig 3428
ReservedCapacityOffering 3429
ReservedCapacitySummary 3433
ResolvedAttributes 3437
ResourceCatalog 3438
ResourceConfig 3440
ResourceConfigForUpdate 3445
ResourceLimits 3446
ResourceSharingConfig 3448
ResourceSpec 3450
RetentionPolicy 3454
RetryStrategy 3455
RoleGroupAssignment 3456
RollingDeploymentPolicy 3458
RollingUpdatePolicy 3459
RSessionAppSettings 3461
RStudioServerProAppSettings 3463
RStudioServerProDomainSettings 3464
RStudioServerProDomainSettingsForUpdate 3466
S3DataSource 3468
S3FileSystem 3473

S3FileSystemConfig 3474
S3ModelDataSource 3475
S3Presign 3479
S3StorageConfig 3480
ScalingPolicy 3482
ScalingPolicyMetric 3483
ScalingPolicyObjective 3484
ScheduleConfig 3485
ScheduledUpdateConfig 3488
SchedulerConfig 3489
SearchExpression 3491
SearchRecord 3493
SecondaryStatusTransition 3497
SelectedStep 3500
SelectiveExecutionConfig 3501
SelectiveExecutionResult 3503
ServerlessJobConfig 3504
ServiceCatalogProvisionedProductDetails 3507
ServiceCatalogProvisioningDetails 3509
ServiceCatalogProvisioningUpdateDetails 3511
SessionChainingConfig 3512
ShadowModeConfig 3513
ShadowModelVariantConfig 3514
SharingSettings 3515
ShuffleConfig 3517
SourceAlgorithm 3518
SourceAlgorithmSpecification 3520
SourceIpConfig 3521
SpaceAppLifecycleManagement 3523
SpaceCodeEditorAppSettings 3524
SpaceDetails 3526
SpaceIdleSettings 3529
SpaceJupyterLabAppSettings 3530
SpaceSettings 3532
SpaceSettingsSummary 3535
SpaceSharingSettings 3537

SpaceSharingSettingsSummary 3538
SpaceStorageSettings 3539
Stairs 3540
StoppingCondition 3542
StudioLifecycleConfigDetails 3545
StudioWebPortalSettings 3547
SubscribedWorkteam 3550
SuggestionQuery 3552
TabularJobConfig 3553
TabularResolvedAttributes 3557
Tag 3558
TargetPlatform 3560
TargetTrackingScalingPolicyConfiguration 3562
TemplateProviderDetail 3563
TensorBoardAppSettings 3564
TensorBoardOutputConfig 3565
TextClassificationJobConfig 3566
TextGenerationJobConfig 3568
TextGenerationResolvedAttributes 3571
ThroughputConfig 3572
ThroughputConfigDescription 3574
ThroughputConfigUpdate 3576
TimeSeriesConfig 3578
TimeSeriesForecastingJobConfig 3580
TimeSeriesForecastingSettings 3584
TimeSeriesTransformations 3586
TotalHits 3588
TrackingServerSummary 3590
TrafficPattern 3593
TrafficRoutingConfig 3595
TrainingImageConfig 3597
TrainingJob 3598
TrainingJobDefinition 3610
TrainingJobStatusCounters 3613
TrainingJobStepMetadata 3615
TrainingJobSummary 3616

TrainingPlanExtension 3619
TrainingPlanExtensionOffering 3622
TrainingPlanFilter 3624
TrainingPlanOffering 3625
TrainingPlanSummary 3628
TrainingProgressInfo 3633
TrainingRepositoryAuthConfig 3635
TrainingSpecification 3636
TransformDataSource 3640
TransformInput 3641
TransformJob 3644
TransformJobDefinition 3651
TransformJobStepMetadata 3654
TransformJobSummary 3655
TransformOutput 3657
TransformResources 3660
TransformS3DataSource 3663
Trial 3665
TrialComponent 3668
TrialComponentArtifact 3673
TrialComponentMetricSummary 3675
TrialComponentParameterValue 3678
TrialComponentSimpleSummary 3679
TrialComponentSource 3681
TrialComponentSourceDetail 3682
TrialComponentStatus 3684
TrialComponentSummary 3685
TrialSource 3688
TrialSummary 3689
TrustedIdentityPropagationSettings 3691
TtlDuration 3692
TuningJobCompletionCriteria 3693
TuningJobStepMetaData 3694
UiConfig 3695
UiTemplate 3698
UiTemplateInfo 3699

UltraServer ... 3700
UltraServerInfo ... 3704
UltraServerSummary ... 3705
UnifiedStudioSettings ... 3707
UpdateClusterSoftwareInstanceGroupSpecification ... 3710
UpdateTemplateProvider ... 3711
USD ... 3712
UserContext ... 3714
UserProfileDetails ... 3716
UserSettings ... 3718
VariantProperty ... 3724
VectorConfig ... 3725
Vertex ... 3726
VisibilityConditions ... 3728
VpcConfig ... 3730
WarmPoolStatus ... 3732
WorkerAccessConfiguration ... 3734
Workforce ... 3735
WorkforceVpcConfigRequest ... 3739
WorkforceVpcConfigResponse ... 3741
WorkloadSpec ... 3743
WorkspaceSettings ... 3744
Workteam ... 3746
Amazon SageMaker Runtime ... 3748
PayloadPart ... 3749
ResponseStream ... 3750
Amazon Sagemaker Edge Manager ... 3751
Checksum ... 3752
Definition ... 3753
DeploymentModel ... 3755
DeploymentResult ... 3758
EdgeDeployment ... 3760
EdgeMetric ... 3762
Model ... 3764
Amazon SageMaker Feature Store Runtime ... 3765
BatchGetRecordError ... 3766

BatchGetRecordIdentifier ... 3768
BatchGetRecordResultDetail ... 3770
FeatureValue ... 3772
TtlDuration ... 3774
Amazon SageMaker geospatial capabilities ... 3774
AreaOfInterest ... 3777
AreaOfInterestGeometry ... 3778
AssetValue ... 3779
BandMathConfigInput ... 3780
CloudMaskingConfigInput ... 3781
CloudRemovalConfigInput ... 3782
CustomIndicesInput ... 3784
EarthObservationJobErrorDetails ... 3785
EoCloudCoverInput ... 3786
ExportErrorDetails ... 3787
ExportErrorDetailsOutput ... 3788
ExportS3DataInput ... 3789
ExportVectorEnrichmentJobOutputConfig ... 3790
Filter ... 3791
Geometry ... 3793
GeoMosaicConfigInput ... 3794
InputConfigInput ... 3795
InputConfigOutput ... 3796
ItemSource ... 3797
JobConfigInput ... 3799
LandCoverSegmentationConfigInput ... 3802
LandsatCloudCoverLandInput ... 3803
ListEarthObservationJobOutputConfig ... 3804
ListVectorEnrichmentJobOutputConfig ... 3806
MapMatchingConfig ... 3808
MultiPolygonGeometryInput ... 3810
Operation ... 3811
OutputBand ... 3812
OutputConfigInput ... 3813
OutputResolutionResamplingInput ... 3814
OutputResolutionStackInput ... 3815

PlatformInput ... 3816
PolygonGeometryInput ... 3817
Properties ... 3818
Property ... 3820
PropertyFilter ... 3822
PropertyFilters ... 3823
RasterDataCollectionMetadata ... 3824
RasterDataCollectionQueryInput ... 3826
RasterDataCollectionQueryOutput ... 3828
RasterDataCollectionQueryWithBandFilterInput ... 3830
ResamplingConfigInput ... 3832
ReverseGeocodingConfig ... 3834
StackConfigInput ... 3835
TemporalStatisticsConfigInput ... 3836
TimeRangeFilterInput ... 3838
TimeRangeFilterOutput ... 3839
UserDefined ... 3840
VectorEnrichmentJobConfig ... 3841
VectorEnrichmentJobDataSourceConfigInput ... 3842
VectorEnrichmentJobErrorDetails ... 3843
VectorEnrichmentJobExportErrorDetails ... 3844
VectorEnrichmentJobInputConfig ... 3845
VectorEnrichmentJobS3Data ... 3846
ViewOffNadirInput ... 3847
ViewSunAzimuthInput ... 3848
ViewSunElevationInput ... 3849
ZonalStatisticsConfigInput ... 3850
Amazon SageMaker Metrics Service ... 3851
BatchPutMetricsError ... 3852
MetricQuery ... 3853
MetricQueryResult ... 3855
RawMetricData ... 3857
Amazon SageMaker Job Runtime Service ... 3858
Common Parameters ... 3859
Common Error Types ... 3862

Welcome

Amazon SageMaker Service

Provides APIs for creating and managing SageMaker resources.

Other Resources:

- SageMaker Developer Guide
- Amazon Augmented AI Runtime API Reference

Amazon SageMaker Runtime

The Amazon SageMaker AI runtime API.

Amazon Sagemaker Edge Manager

SageMaker Edge Manager dataplane service for communicating with active agents.

Amazon SageMaker Feature Store Runtime

Contains all data plane API operations and data types for the Amazon SageMaker Feature Store. Use this API to put, delete, and retrieve (get) features from a feature store.

Use the following operations to configure your `OnlineStore` and `OfflineStore` features, and to create and manage feature groups:

- CreateFeatureGroup
- DeleteFeatureGroup
- DescribeFeatureGroup
- ListFeatureGroups

Amazon SageMaker geospatial capabilities

Provides APIs for creating and managing SageMaker geospatial resources.

Amazon SageMaker Metrics Service

Contains all data plane API operations and data types for Amazon SageMaker Metrics. Use these APIs to put and retrieve (get) features related to your training run.

- BatchPutMetrics

Amazon SageMaker Job Runtime Service

Amazon SageMaker Job Runtime provides APIs that enable communication between your agent and the policy model during multi-turn reinforcement learning jobs. These APIs support inference requests to the policy model, rollout lifecycle management, and reward submission.

Actions

The following actions are supported by Amazon SageMaker Service:

- AddAssociation
- AddTags
- AssociateTrialComponent
- AttachClusterNodeVolume
- BatchAddClusterNodes
- BatchDeleteClusterNodes
- BatchDescribeModelPackage
- BatchRebootClusterNodes
- BatchReplaceClusterNodes
- CreateAction
- CreateAIBenchmarkJob
- CreateAIRecommendationJob
- CreateAIWorkloadConfig
- CreateAlgorithm
- CreateApp
- CreateAppImageConfig
- CreateArtifact
- CreateAutoMLJob
- CreateAutoMLJobV2
- CreateCluster
- CreateClusterSchedulerConfig
- CreateCodeRepository
- CreateCompilationJob
- CreateComputeQuota
- CreateContext
- CreateDataQualityJobDefinition
- CreateDeviceFleet

- CreateDomain
- CreateEdgeDeploymentPlan
- CreateEdgeDeploymentStage
- CreateEdgePackagingJob
- CreateEndpoint
- CreateEndpointConfig
- CreateExperiment
- CreateFeatureGroup
- CreateFlowDefinition
- CreateHub
- CreateHubContentPresignedUrls
- CreateHubContentReference
- CreateHumanTaskUi
- CreateHyperParameterTuningJob
- CreateImage
- CreateImageVersion
- CreateInferenceComponent
- CreateInferenceExperiment
- CreateInferenceRecommendationsJob
- CreateJob
- CreateLabelingJob
- CreateMlflowApp
- CreateMlflowTrackingServer
- CreateModel
- CreateModelBiasJobDefinition
- CreateModelCard
- CreateModelCardExportJob
- CreateModelExplainabilityJobDefinition
- CreateModelPackage
- CreateModelPackageGroup

- CreateModelQualityJobDefinition
- CreateMonitoringSchedule
- CreateNotebookInstance
- CreateNotebookInstanceLifecycleConfig
- CreateOptimizationJob
- CreatePartnerApp
- CreatePartnerAppPresignedUrl
- CreatePipeline
- CreatePresignedDomainUrl
- CreatePresignedMlflowAppUrl
- CreatePresignedMlflowTrackingServerUrl
- CreatePresignedNotebookInstanceUrl
- CreateProcessingJob
- CreateProject
- CreateSpace
- CreateStudioLifecycleConfig
- CreateTrainingJob
- CreateTrainingPlan
- CreateTransformJob
- CreateTrial
- CreateTrialComponent
- CreateUserProfile
- CreateWorkforce
- CreateWorkteam
- DeleteAction
- DeleteAIBenchmarkJob
- DeleteAIRecommendationJob
- DeleteAIWorkloadConfig
- DeleteAlgorithm
- DeleteApp

- DeleteAppImageConfig
- DeleteArtifact
- DeleteAssociation
- DeleteCluster
- DeleteClusterSchedulerConfig
- DeleteCodeRepository
- DeleteCompilationJob
- DeleteComputeQuota
- DeleteContext
- DeleteDataQualityJobDefinition
- DeleteDeviceFleet
- DeleteDomain
- DeleteEdgeDeploymentPlan
- DeleteEdgeDeploymentStage
- DeleteEndpoint
- DeleteEndpointConfig
- DeleteExperiment
- DeleteFeatureGroup
- DeleteFlowDefinition
- DeleteHub
- DeleteHubContent
- DeleteHubContentReference
- DeleteHumanTaskUi
- DeleteHyperParameterTuningJob
- DeleteImage
- DeleteImageVersion
- DeleteInferenceComponent
- DeleteInferenceExperiment
- DeleteJob
- DeleteMlflowApp

- DeleteMlflowTrackingServer
- DeleteModel
- DeleteModelBiasJobDefinition
- DeleteModelCard
- DeleteModelExplainabilityJobDefinition
- DeleteModelPackage
- DeleteModelPackageGroup
- DeleteModelPackageGroupPolicy
- DeleteModelQualityJobDefinition
- DeleteMonitoringSchedule
- DeleteNotebookInstance
- DeleteNotebookInstanceLifecycleConfig
- DeleteOptimizationJob
- DeletePartnerApp
- DeletePipeline
- DeleteProcessingJob
- DeleteProject
- DeleteSpace
- DeleteStudioLifecycleConfig
- DeleteTags
- DeleteTrainingJob
- DeleteTrial
- DeleteTrialComponent
- DeleteUserProfile
- DeleteWorkforce
- DeleteWorkteam
- DeregisterDevices
- DescribeAction
- DescribeAIBenchmarkJob
- DescribeAIRecommendationJob

- DescribeAIWorkloadConfig
- DescribeAlgorithm
- DescribeApp
- DescribeAppImageConfig
- DescribeArtifact
- DescribeAutoMLJob
- DescribeAutoMLJobV2
- DescribeCluster
- DescribeClusterEvent
- DescribeClusterNode
- DescribeClusterSchedulerConfig
- DescribeCodeRepository
- DescribeCompilationJob
- DescribeComputeQuota
- DescribeContext
- DescribeDataQualityJobDefinition
- DescribeDevice
- DescribeDeviceFleet
- DescribeDomain
- DescribeEdgeDeploymentPlan
- DescribeEdgePackagingJob
- DescribeEndpoint
- DescribeEndpointConfig
- DescribeExperiment
- DescribeFeatureGroup
- DescribeFeatureMetadata
- DescribeFlowDefinition
- DescribeHub
- DescribeHubContent
- DescribeHumanTaskUi

- DescribeHyperParameterTuningJob
- DescribeImage
- DescribeImageVersion
- DescribeInferenceComponent
- DescribeInferenceExperiment
- DescribeInferenceRecommendationsJob
- DescribeJob
- DescribeJobSchemaVersion
- DescribeLabelingJob
- DescribeLineageGroup
- DescribeMlflowApp
- DescribeMlflowTrackingServer
- DescribeModel
- DescribeModelBiasJobDefinition
- DescribeModelCard
- DescribeModelCardExportJob
- DescribeModelExplainabilityJobDefinition
- DescribeModelPackage
- DescribeModelPackageGroup
- DescribeModelQualityJobDefinition
- DescribeMonitoringSchedule
- DescribeNotebookInstance
- DescribeNotebookInstanceLifecycleConfig
- DescribeOptimizationJob
- DescribePartnerApp
- DescribePipeline
- DescribePipelineDefinitionForExecution
- DescribePipelineExecution
- DescribeProcessingJob
- DescribeProject

- DescribeReservedCapacity
- DescribeSpace
- DescribeStudioLifecycleConfig
- DescribeSubscribedWorkteam
- DescribeTrainingJob
- DescribeTrainingPlan
- DescribeTrainingPlanExtensionHistory
- DescribeTransformJob
- DescribeTrial
- DescribeTrialComponent
- DescribeUserProfile
- DescribeWorkforce
- DescribeWorkteam
- DetachClusterNodeVolume
- DisableSagemakerServicecatalogPortfolio
- DisassociateTrialComponent
- EnableSagemakerServicecatalogPortfolio
- ExtendTrainingPlan
- GetDeviceFleetReport
- GetLineageGroupPolicy
- GetModelPackageGroupPolicy
- GetSagemakerServicecatalogPortfolioStatus
- GetScalingConfigurationRecommendation
- GetSearchSuggestions
- ImportHubContent
- ListActions
- ListAIBenchmarkJobs
- ListAIRecommendationJobs
- ListAIWorkloadConfigs
- ListAlgorithms

- ListAliases
- ListAppImageConfigs
- ListApps
- ListArtifacts
- ListAssociations
- ListAutoMLJobs
- ListCandidatesForAutoMLJob
- ListClusterEvents
- ListClusterNodes
- ListClusters
- ListClusterSchedulerConfigs
- ListCodeRepositories
- ListCompilationJobs
- ListComputeQuotas
- ListContexts
- ListDataQualityJobDefinitions
- ListDeviceFleets
- ListDevices
- ListDomains
- ListEdgeDeploymentPlans
- ListEdgePackagingJobs
- ListEndpointConfigs
- ListEndpoints
- ListExperiments
- ListFeatureGroups
- ListFlowDefinitions
- ListHubContents
- ListHubContentVersions
- ListHubs
- ListHumanTaskUis

- ListHyperParameterTuningJobs
- ListImages
- ListImageVersions
- ListInferenceComponents
- ListInferenceExperiments
- ListInferenceRecommendationsJobs
- ListInferenceRecommendationsJobSteps
- ListJobs
- ListJobSchemaVersions
- ListLabelingJobs
- ListLabelingJobsForWorkteam
- ListLineageGroups
- ListMlflowApps
- ListMlflowTrackingServers
- ListModelBiasJobDefinitions
- ListModelCardExportJobs
- ListModelCards
- ListModelCardVersions
- ListModelExplainabilityJobDefinitions
- ListModelMetadata
- ListModelPackageGroups
- ListModelPackages
- ListModelQualityJobDefinitions
- ListModels
- ListMonitoringAlertHistory
- ListMonitoringAlerts
- ListMonitoringExecutions
- ListMonitoringSchedules
- ListNotebookInstanceLifecycleConfigs
- ListNotebookInstances

- ListOptimizationJobs
- ListPartnerApps
- ListPipelineExecutions
- ListPipelineExecutionSteps
- ListPipelineParametersForExecution
- ListPipelines
- ListPipelineVersions
- ListProcessingJobs
- ListProjects
- ListResourceCatalogs
- ListSpaces
- ListStageDevices
- ListStudioLifecycleConfigs
- ListSubscribedWorkteams
- ListTags
- ListTrainingJobs
- ListTrainingJobsForHyperParameterTuningJob
- ListTrainingPlans
- ListTransformJobs
- ListTrialComponents
- ListTrials
- ListUltraServersByReservedCapacity
- ListUserProfiles
- ListWorkforces
- ListWorkteams
- PutModelPackageGroupPolicy
- QueryLineage
- RegisterDevices
- RenderUiTemplate
- RetryPipelineExecution

- Search
- SearchTrainingPlanOfferings
- SendPipelineExecutionStepFailure
- SendPipelineExecutionStepSuccess
- StartClusterHealthCheck
- StartEdgeDeploymentStage
- StartInferenceExperiment
- StartMlflowTrackingServer
- StartMonitoringSchedule
- StartNotebookInstance
- StartPipelineExecution
- StartSession
- StopAIBenchmarkJob
- StopAIRecommendationJob
- StopAutoMLJob
- StopCompilationJob
- StopEdgeDeploymentStage
- StopEdgePackagingJob
- StopHyperParameterTuningJob
- StopInferenceExperiment
- StopInferenceRecommendationsJob
- StopJob
- StopLabelingJob
- StopMlflowTrackingServer
- StopMonitoringSchedule
- StopNotebookInstance
- StopOptimizationJob
- StopPipelineExecution
- StopProcessingJob
- StopTrainingJob

- StopTransformJob
- UpdateAction
- UpdateAppImageConfig
- UpdateArtifact
- UpdateCluster
- UpdateClusterSchedulerConfig
- UpdateClusterSoftware
- UpdateCodeRepository
- UpdateComputeQuota
- UpdateContext
- UpdateDeviceFleet
- UpdateDevices
- UpdateDomain
- UpdateEndpoint
- UpdateEndpointWeightsAndCapacities
- UpdateExperiment
- UpdateFeatureGroup
- UpdateFeatureMetadata
- UpdateHub
- UpdateHubContent
- UpdateHubContentReference
- UpdateImage
- UpdateImageVersion
- UpdateInferenceComponent
- UpdateInferenceComponentRuntimeConfig
- UpdateInferenceExperiment
- UpdateMlflowApp
- UpdateMlflowTrackingServer
- UpdateModelCard
- UpdateModelPackage

- UpdateMonitoringAlert
- UpdateMonitoringSchedule
- UpdateNotebookInstance
- UpdateNotebookInstanceLifecycleConfig
- UpdatePartnerApp
- UpdatePipeline
- UpdatePipelineExecution
- UpdatePipelineVersion
- UpdateProject
- UpdateSpace
- UpdateTrainingJob
- UpdateTrial
- UpdateTrialComponent
- UpdateUserProfile
- UpdateWorkforce
- UpdateWorkteam

The following actions are supported by Amazon SageMaker Runtime:

- InvokeEndpoint
- InvokeEndpointAsync
- InvokeEndpointWithResponseStream

The following actions are supported by Amazon Sagemaker Edge Manager:

- GetDeployments
- GetDeviceRegistration
- SendHeartbeat

The following actions are supported by Amazon SageMaker Feature Store Runtime:

- BatchGetRecord

- DeleteRecord
- GetRecord
- PutRecord

The following actions are supported by Amazon SageMaker geospatial capabilities:

- DeleteEarthObservationJob
- DeleteVectorEnrichmentJob
- ExportEarthObservationJob
- ExportVectorEnrichmentJob
- GetEarthObservationJob
- GetRasterDataCollection
- GetTile
- GetVectorEnrichmentJob
- ListEarthObservationJobs
- ListRasterDataCollections
- ListTagsForResource
- ListVectorEnrichmentJobs
- SearchRasterDataCollection
- StartEarthObservationJob
- StartVectorEnrichmentJob
- StopEarthObservationJob
- StopVectorEnrichmentJob
- TagResource
- UntagResource

The following actions are supported by Amazon SageMaker Metrics Service:

- BatchGetMetrics
- BatchPutMetrics

The following actions are supported by Amazon SageMaker Job Runtime Service:

- CompleteRollout
- Sample
- SampleWithResponseStream
- UpdateReward

Amazon SageMaker Service

The following actions are supported by Amazon SageMaker Service:

- AddAssociation
- AddTags
- AssociateTrialComponent
- AttachClusterNodeVolume
- BatchAddClusterNodes
- BatchDeleteClusterNodes
- BatchDescribeModelPackage
- BatchRebootClusterNodes
- BatchReplaceClusterNodes
- CreateAction
- CreateAIBenchmarkJob
- CreateAIRecommendationJob
- CreateAIWorkloadConfig
- CreateAlgorithm
- CreateApp
- CreateAppImageConfig
- CreateArtifact
- CreateAutoMLJob
- CreateAutoMLJobV2
- CreateCluster
- CreateClusterSchedulerConfig
- CreateCodeRepository

- CreateCompilationJob
- CreateComputeQuota
- CreateContext
- CreateDataQualityJobDefinition
- CreateDeviceFleet
- CreateDomain
- CreateEdgeDeploymentPlan
- CreateEdgeDeploymentStage
- CreateEdgePackagingJob
- CreateEndpoint
- CreateEndpointConfig
- CreateExperiment
- CreateFeatureGroup
- CreateFlowDefinition
- CreateHub
- CreateHubContentPresignedUrls
- CreateHubContentReference
- CreateHumanTaskUi
- CreateHyperParameterTuningJob
- CreateImage
- CreateImageVersion
- CreateInferenceComponent
- CreateInferenceExperiment
- CreateInferenceRecommendationsJob
- CreateJob
- CreateLabelingJob
- CreateMlflowApp
- CreateMlflowTrackingServer
- CreateModel
- CreateModelBiasJobDefinition

- CreateModelCard
- CreateModelCardExportJob
- CreateModelExplainabilityJobDefinition
- CreateModelPackage
- CreateModelPackageGroup
- CreateModelQualityJobDefinition
- CreateMonitoringSchedule
- CreateNotebookInstance
- CreateNotebookInstanceLifecycleConfig
- CreateOptimizationJob
- CreatePartnerApp
- CreatePartnerAppPresignedUrl
- CreatePipeline
- CreatePresignedDomainUrl
- CreatePresignedMlflowAppUrl
- CreatePresignedMlflowTrackingServerUrl
- CreatePresignedNotebookInstanceUrl
- CreateProcessingJob
- CreateProject
- CreateSpace
- CreateStudioLifecycleConfig
- CreateTrainingJob
- CreateTrainingPlan
- CreateTransformJob
- CreateTrial
- CreateTrialComponent
- CreateUserProfile
- CreateWorkforce
- CreateWorkteam
- DeleteAction

- DeleteAIBenchmarkJob
- DeleteAIRecommendationJob
- DeleteAIWorkloadConfig
- DeleteAlgorithm
- DeleteApp
- DeleteAppImageConfig
- DeleteArtifact
- DeleteAssociation
- DeleteCluster
- DeleteClusterSchedulerConfig
- DeleteCodeRepository
- DeleteCompilationJob
- DeleteComputeQuota
- DeleteContext
- DeleteDataQualityJobDefinition
- DeleteDeviceFleet
- DeleteDomain
- DeleteEdgeDeploymentPlan
- DeleteEdgeDeploymentStage
- DeleteEndpoint
- DeleteEndpointConfig
- DeleteExperiment
- DeleteFeatureGroup
- DeleteFlowDefinition
- DeleteHub
- DeleteHubContent
- DeleteHubContentReference
- DeleteHumanTaskUi
- DeleteHyperParameterTuningJob
- DeleteImage

- DeleteImageVersion
- DeleteInferenceComponent
- DeleteInferenceExperiment
- DeleteJob
- DeleteMlflowApp
- DeleteMlflowTrackingServer
- DeleteModel
- DeleteModelBiasJobDefinition
- DeleteModelCard
- DeleteModelExplainabilityJobDefinition
- DeleteModelPackage
- DeleteModelPackageGroup
- DeleteModelPackageGroupPolicy
- DeleteModelQualityJobDefinition
- DeleteMonitoringSchedule
- DeleteNotebookInstance
- DeleteNotebookInstanceLifecycleConfig
- DeleteOptimizationJob
- DeletePartnerApp
- DeletePipeline
- DeleteProcessingJob
- DeleteProject
- DeleteSpace
- DeleteStudioLifecycleConfig
- DeleteTags
- DeleteTrainingJob
- DeleteTrial
- DeleteTrialComponent
- DeleteUserProfile
- DeleteWorkforce

- DeleteWorkteam
- DeregisterDevices
- DescribeAction
- DescribeAIBenchmarkJob
- DescribeAIRecommendationJob
- DescribeAIWorkloadConfig
- DescribeAlgorithm
- DescribeApp
- DescribeAppImageConfig
- DescribeArtifact
- DescribeAutoMLJob
- DescribeAutoMLJobV2
- DescribeCluster
- DescribeClusterEvent
- DescribeClusterNode
- DescribeClusterSchedulerConfig
- DescribeCodeRepository
- DescribeCompilationJob
- DescribeComputeQuota
- DescribeContext
- DescribeDataQualityJobDefinition
- DescribeDevice
- DescribeDeviceFleet
- DescribeDomain
- DescribeEdgeDeploymentPlan
- DescribeEdgePackagingJob
- DescribeEndpoint
- DescribeEndpointConfig
- DescribeExperiment
- DescribeFeatureGroup

- DescribeFeatureMetadata
- DescribeFlowDefinition
- DescribeHub
- DescribeHubContent
- DescribeHumanTaskUi
- DescribeHyperParameterTuningJob
- DescribeImage
- DescribeImageVersion
- DescribeInferenceComponent
- DescribeInferenceExperiment
- DescribeInferenceRecommendationsJob
- DescribeJob
- DescribeJobSchemaVersion
- DescribeLabelingJob
- DescribeLineageGroup
- DescribeMlflowApp
- DescribeMlflowTrackingServer
- DescribeModel
- DescribeModelBiasJobDefinition
- DescribeModelCard
- DescribeModelCardExportJob
- DescribeModelExplainabilityJobDefinition
- DescribeModelPackage
- DescribeModelPackageGroup
- DescribeModelQualityJobDefinition
- DescribeMonitoringSchedule
- DescribeNotebookInstance
- DescribeNotebookInstanceLifecycleConfig
- DescribeOptimizationJob
- DescribePartnerApp

- DescribePipeline
- DescribePipelineDefinitionForExecution
- DescribePipelineExecution
- DescribeProcessingJob
- DescribeProject
- DescribeReservedCapacity
- DescribeSpace
- DescribeStudioLifecycleConfig
- DescribeSubscribedWorkteam
- DescribeTrainingJob
- DescribeTrainingPlan
- DescribeTrainingPlanExtensionHistory
- DescribeTransformJob
- DescribeTrial
- DescribeTrialComponent
- DescribeUserProfile
- DescribeWorkforce
- DescribeWorkteam
- DetachClusterNodeVolume
- DisableSagemakerServicecatalogPortfolio
- DisassociateTrialComponent
- EnableSagemakerServicecatalogPortfolio
- ExtendTrainingPlan
- GetDeviceFleetReport
- GetLineageGroupPolicy
- GetModelPackageGroupPolicy
- GetSagemakerServicecatalogPortfolioStatus
- GetScalingConfigurationRecommendation
- GetSearchSuggestions
- ImportHubContent

- ListActions
- ListAIBenchmarkJobs
- ListAIRecommendationJobs
- ListAIWorkloadConfigs
- ListAlgorithms
- ListAliases
- ListAppImageConfigs
- ListApps
- ListArtifacts
- ListAssociations
- ListAutoMLJobs
- ListCandidatesForAutoMLJob
- ListClusterEvents
- ListClusterNodes
- ListClusters
- ListClusterSchedulerConfigs
- ListCodeRepositories
- ListCompilationJobs
- ListComputeQuotas
- ListContexts
- ListDataQualityJobDefinitions
- ListDeviceFleets
- ListDevices
- ListDomains
- ListEdgeDeploymentPlans
- ListEdgePackagingJobs
- ListEndpointConfigs
- ListEndpoints
- ListExperiments
- ListFeatureGroups

- ListFlowDefinitions
- ListHubContents
- ListHubContentVersions
- ListHubs
- ListHumanTaskUis
- ListHyperParameterTuningJobs
- ListImages
- ListImageVersions
- ListInferenceComponents
- ListInferenceExperiments
- ListInferenceRecommendationsJobs
- ListInferenceRecommendationsJobSteps
- ListJobs
- ListJobSchemaVersions
- ListLabelingJobs
- ListLabelingJobsForWorkteam
- ListLineageGroups
- ListMlflowApps
- ListMlflowTrackingServers
- ListModelBiasJobDefinitions
- ListModelCardExportJobs
- ListModelCards
- ListModelCardVersions
- ListModelExplainabilityJobDefinitions
- ListModelMetadata
- ListModelPackageGroups
- ListModelPackages
- ListModelQualityJobDefinitions
- ListModels
- ListMonitoringAlertHistory

- ListMonitoringAlerts
- ListMonitoringExecutions
- ListMonitoringSchedules
- ListNotebookInstanceLifecycleConfigs
- ListNotebookInstances
- ListOptimizationJobs
- ListPartnerApps
- ListPipelineExecutions
- ListPipelineExecutionSteps
- ListPipelineParametersForExecution
- ListPipelines
- ListPipelineVersions
- ListProcessingJobs
- ListProjects
- ListResourceCatalogs
- ListSpaces
- ListStageDevices
- ListStudioLifecycleConfigs
- ListSubscribedWorkteams
- ListTags
- ListTrainingJobs
- ListTrainingJobsForHyperParameterTuningJob
- ListTrainingPlans
- ListTransformJobs
- ListTrialComponents
- ListTrials
- ListUltraServersByReservedCapacity
- ListUserProfiles
- ListWorkforces
- ListWorkteams

- PutModelPackageGroupPolicy
- QueryLineage
- RegisterDevices
- RenderUiTemplate
- RetryPipelineExecution
- Search
- SearchTrainingPlanOfferings
- SendPipelineExecutionStepFailure
- SendPipelineExecutionStepSuccess
- StartClusterHealthCheck
- StartEdgeDeploymentStage
- StartInferenceExperiment
- StartMlflowTrackingServer
- StartMonitoringSchedule
- StartNotebookInstance
- StartPipelineExecution
- StartSession
- StopAIBenchmarkJob
- StopAIRecommendationJob
- StopAutoMLJob
- StopCompilationJob
- StopEdgeDeploymentStage
- StopEdgePackagingJob
- StopHyperParameterTuningJob
- StopInferenceExperiment
- StopInferenceRecommendationsJob
- StopJob
- StopLabelingJob
- StopMlflowTrackingServer
- StopMonitoringSchedule

- StopNotebookInstance
- StopOptimizationJob
- StopPipelineExecution
- StopProcessingJob
- StopTrainingJob
- StopTransformJob
- UpdateAction
- UpdateAppImageConfig
- UpdateArtifact
- UpdateCluster
- UpdateClusterSchedulerConfig
- UpdateClusterSoftware
- UpdateCodeRepository
- UpdateComputeQuota
- UpdateContext
- UpdateDeviceFleet
- UpdateDevices
- UpdateDomain
- UpdateEndpoint
- UpdateEndpointWeightsAndCapacities
- UpdateExperiment
- UpdateFeatureGroup
- UpdateFeatureMetadata
- UpdateHub
- UpdateHubContent
- UpdateHubContentReference
- UpdateImage
- UpdateImageVersion
- UpdateInferenceComponent
- UpdateInferenceComponentRuntimeConfig

- UpdateInferenceExperiment
- UpdateMlflowApp
- UpdateMlflowTrackingServer
- UpdateModelCard
- UpdateModelPackage
- UpdateMonitoringAlert
- UpdateMonitoringSchedule
- UpdateNotebookInstance
- UpdateNotebookInstanceLifecycleConfig
- UpdatePartnerApp
- UpdatePipeline
- UpdatePipelineExecution
- UpdatePipelineVersion
- UpdateProject
- UpdateSpace
- UpdateTrainingJob
- UpdateTrial
- UpdateTrialComponent
- UpdateUserProfile
- UpdateWorkforce
- UpdateWorkteam

AddAssociation

Service: Amazon SageMaker Service

Creates an *association* between the source and the destination. A source can be associated with multiple destinations, and a destination can be associated with multiple sources. An association is a lineage tracking entity. For more information, see Amazon SageMaker ML Lineage Tracking.

Request Syntax

```
{
   "AssociationType": "string",
   "DestinationArn": "string",
   "SourceArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AssociationType

The type of association. The following are suggested uses for each type. Amazon SageMaker places no restrictions on their use.

- ContributedTo - The source contributed to the destination or had a part in enabling the destination. For example, the training data contributed to the training job.
- AssociatedWith - The source is connected to the destination. For example, an approval workflow is associated with a model deployment.
- DerivedFrom - The destination is a modification of the source. For example, a digest output of a channel input for a processing job is derived from the original inputs.
- Produced - The source generated the destination. For example, a training job produced a model artifact.

Type: String

Valid Values: `ContributedTo | AssociatedWith | DerivedFrom | Produced | SameAs`

Required: No

DestinationArn

The Amazon Resource Name (ARN) of the destination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: Yes

SourceArn

The ARN of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: Yes

Response Syntax

```
{
   "DestinationArn": "string",
   "SourceArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

DestinationArn

The Amazon Resource Name (ARN) of the destination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

SourceArn

The ARN of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

AddTags

Service: Amazon SageMaker Service

Adds or overwrites one or more tags for the specified SageMaker resource. You can add tags to notebook instances, training jobs, hyperparameter tuning jobs, batch transform jobs, models, labeling jobs, work teams, endpoint configurations, and endpoints.

Each tag consists of a key and an optional value. Tag keys must be unique per resource. For more information about tags, see For more information, see AWS Tagging Strategies.

Note

Tags that you add to a hyperparameter tuning job by calling this API are also added to any training jobs that the hyperparameter tuning job launches after you call this API, but not to training jobs that the hyperparameter tuning job launched before you called this API. To make sure that the tags associated with a hyperparameter tuning job are also added to all training jobs that the hyperparameter tuning job launches, add the tags when you first create the tuning job by specifying them in the `Tags` parameter of CreateHyperParameterTuningJob

Note

Tags that you add to a SageMaker Domain or User Profile by calling this API are also added to any Apps that the Domain or User Profile launches after you call this API, but not to Apps that the Domain or User Profile launched before you called this API. To make sure that the tags associated with a Domain or User Profile are also added to all Apps that the Domain or User Profile launches, add the tags when you first create the Domain or User Profile by specifying them in the `Tags` parameter of CreateDomain or CreateUserProfile.

Request Syntax

```
{
   "ResourceArn": "string",
   "Tags": [
      {
         "Key": "string",
```

```
            "Value": "string"
        }
    ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ResourceArn

The Amazon Resource Name (ARN) of the resource that you want to tag.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z-]*:sagemaker:[a-z0-9-]*:[0-9]{12}:.+`

Required: Yes

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: Yes

Response Syntax

```
{
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
```

```
    ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Tags

A list of tags associated with the SageMaker resource.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

AssociateTrialComponent

Service: Amazon SageMaker Service

Associates a trial component with a trial. A trial component can be associated with multiple trials. To disassociate a trial component from a trial, call the DisassociateTrialComponent API.

Request Syntax

```
{
   "TrialComponentName": "string",
   "TrialName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrialComponentName

The name of the component to associated with the trial.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

TrialName

The name of the trial to associate with.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "TrialArn": "string",
   "TrialComponentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrialArn

The Amazon Resource Name (ARN) of the trial.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial/.*`

TrialComponentArn

The Amazon Resource Name (ARN) of the trial component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial-component/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

AttachClusterNodeVolume

Service: Amazon SageMaker Service

Attaches your Amazon Elastic Block Store (Amazon EBS) volume to a node in your EKS orchestrated HyperPod cluster.

This API works with the Amazon Elastic Block Store (Amazon EBS) Container Storage Interface (CSI) driver to manage the lifecycle of persistent storage in your HyperPod EKS clusters.

Request Syntax

```
{
   "ClusterArn": "string",
   "NodeId": "string",
   "VolumeId": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterArn

The Amazon Resource Name (ARN) of your SageMaker HyperPod cluster containing the target node. Your cluster must use EKS as the orchestration and be in the `InService` state.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Required: Yes

NodeId

The unique identifier of the cluster node to which you want to attach the volume. The node must belong to your specified HyperPod cluster and cannot be part of a Restricted Instance Group (RIG).

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: Yes

VolumeId

The unique identifier of your EBS volume to attach. The volume must be in the `available` state.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `vol-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: Yes

Response Syntax

```
{
   "AttachTime": number,
   "ClusterArn": "string",
   "DeviceName": "string",
   "NodeId": "string",
   "Status": "string",
   "VolumeId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AttachTime

The timestamp when the volume attachment operation was initiated by the SageMaker HyperPod service.

Type: Timestamp

ClusterArn

The Amazon Resource Name (ARN) of your SageMaker HyperPod cluster where the volume attachment operation was performed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

DeviceName

The device name assigned to your attached volume on the target instance.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `/dev/(xvd[a-z]|sd[b-z]|sd[f-p][1-6])`

NodeId

The unique identifier of the cluster node where your volume was attached.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Status

The current status of your volume attachment operation.

Type: String

Valid Values: `attaching | attached | detaching | detached | busy`

VolumeId

The unique identifier of your EBS volume that was attached.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `vol-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

BatchAddClusterNodes

Service: Amazon SageMaker Service

Adds nodes to a HyperPod cluster by incrementing the target count for one or more instance groups. This operation returns a unique `NodeLogicalId` for each node being added, which can be used to track the provisioning status of the node. This API provides a safer alternative to `UpdateCluster` for scaling operations by avoiding unintended configuration changes.

> **Note**
>
> This API is only supported for clusters using `Continuous` as the `NodeProvisioningMode`.

Request Syntax

```
{
   "ClientToken": "string",
   "ClusterName": "string",
   "NodesToAdd": [
      {
         "AvailabilityZones": [ "string" ],
         "IncrementTargetCountBy": number,
         "InstanceGroupName": "string",
         "InstanceTypes": [ "string" ]
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. This token is valid for 8 hours. If you retry the request with the same client token within this timeframe and the same parameters, the API returns the same set of `NodeLogicalIds` with their latest status.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[\x21-\x7E]+`

Required: No

ClusterName

The name of the HyperPod cluster to which you want to add nodes.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

NodesToAdd

A list of instance groups and the number of nodes to add to each. You can specify up to 5 instance groups in a single request, with a maximum of 50 nodes total across all instance groups.

Type: Array of AddClusterNodeSpecification objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

Response Syntax

```
{
   "Failed": [
      {
         "AvailabilityZones": [ "string" ],
         "ErrorCode": "string",
         "FailedCount": number,
         "InstanceGroupName": "string",
         "InstanceTypes": [ "string" ],
         "Message": "string"
```

```
      }
   ],
   "Successful": [
      {
         "AvailabilityZones": [ "string" ],
         "InstanceGroupName": "string",
         "InstanceTypes": [ "string" ],
         "NodeLogicalId": "string",
         "Status": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Failed

A list of errors that occurred during the node addition operation. Each entry includes the instance group name, error code, number of failed additions, and an error message.

Type: Array of BatchAddClusterNodesError objects

Successful

A list of `NodeLogicalIDs` that were successfully added to the cluster. The `NodeLogicalID` is unique per cluster and does not change between instance replacements. Each entry includes a `NodeLogicalId` that can be used to track the node's provisioning status (with `DescribeClusterNode`), the instance group name, and the current status of the node.

Type: Array of NodeAdditionResult objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

BatchDeleteClusterNodes

Service: Amazon SageMaker Service

Deletes specific nodes within a SageMaker HyperPod cluster. `BatchDeleteClusterNodes` accepts a cluster name and a list of node IDs.

> **Important**
>
> - To safeguard your work, back up your data to Amazon S3 or an FSx for Lustre file system before invoking the API on a worker node group. This will help prevent any potential data loss from the instance root volume. For more information about backup, see Use the backup script provided by SageMaker HyperPod.
> - If you want to invoke this API on an existing cluster, you'll first need to patch the cluster by running the UpdateClusterSoftware API. For more information about patching a cluster, see Update the SageMaker HyperPod platform software of a cluster.

Request Syntax

```
{
   "ClusterName": "string",
   "NodeIds": [ "string" ],
   "NodeLogicalIds": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterName

The name of the SageMaker HyperPod cluster from which to delete the specified nodes.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

NodeIds

A list of node IDs to be deleted from the specified cluster.

> **Note**
>
> - For SageMaker HyperPod clusters using the Slurm workload manager, you cannot remove instances that are configured as Slurm controller nodes.
> - If you need to delete more than 99 instances, contact Support for assistance.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 3000 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: No

NodeLogicalIds

A list of `NodeLogicalIds` identifying the nodes to be deleted. You can specify up to 50 `NodeLogicalIds`. You must specify either `NodeLogicalIds`, `InstanceIds`, or both, with a combined maximum of 50 identifiers.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 99 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Required: No

Response Syntax

```
{
   "Failed": [
      {
         "Code": "string",
         "Message": "string",
         "NodeId": "string"
      }
   ],
   "FailedNodeLogicalIds": [
      {
         "Code": "string",
         "Message": "string",
         "NodeLogicalId": "string"
      }
   ],
   "Successful": [ "string" ],
   "SuccessfulNodeLogicalIds": [ "string" ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Failed

A list of errors encountered when deleting the specified nodes.

Type: Array of BatchDeleteClusterNodesError objects

Array Members: Minimum number of 1 item. Maximum number of 3000 items.

FailedNodeLogicalIds

A list of `NodeLogicalIds` that could not be deleted, along with error information explaining why the deletion failed.

Type: Array of BatchDeleteClusterNodeLogicalIdsError objects

Array Members: Minimum number of 1 item. Maximum number of 99 items.

Successful

A list of node IDs that were successfully deleted from the specified cluster.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 3000 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

SuccessfulNodeLogicalIds

A list of `NodeLogicalIds` that were successfully deleted from the cluster.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 99 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

BatchDescribeModelPackage

Service: Amazon SageMaker Service

This action batch describes a list of versioned model packages

Request Syntax

```
{
   "ModelPackageArnList": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelPackageArnList

The list of Amazon Resource Name (ARN) of the model package groups.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: Yes

Response Syntax

```
{
   "BatchDescribeModelPackageErrorMap": {
      "string" : {
         "ErrorCode": "string",
         "ErrorResponse": "string"
      }
   },
   "ModelPackageSummaries": {
```

```
      "string" : {
         "CreationTime": number,
         "InferenceSpecification": {
            "Containers": [
               {
                  "AdditionalModelDataSources": [
                     {
                        "ChannelName": "string",
                        "S3DataSource": {
                           "CompressionType": "string",
                           "ETag": "string",
                           "HubAccessConfig": {
                              "HubContentArn": "string"
                           },
                           "ManifestEtag": "string",
                           "ManifestS3Uri": "string",
                           "ModelAccessConfig": {
                              "AcceptEula": boolean
                           },
                           "S3DataType": "string",
                           "S3Uri": "string"
                        }
                     }
                  ],
                  "AdditionalS3DataSource": {
                     "CompressionType": "string",
                     "ETag": "string",
                     "S3DataType": "string",
                     "S3Uri": "string"
                  },
                  "BaseModel": {
                     "HubContentName": "string",
                     "HubContentVersion": "string",
                     "RecipeName": "string"
                  },
                  "ContainerHostname": "string",
                  "Environment": {
                     "string" : "string"
                  },
                  "Framework": "string",
                  "FrameworkVersion": "string",
                  "Image": "string",
                  "ImageDigest": "string",
                  "IsCheckpoint": boolean,
```

```
                  "ModelDataETag": "string",
                  "ModelDataSource": {
                     "S3DataSource": {
                        "CompressionType": "string",
                        "ETag": "string",
                        "HubAccessConfig": {
                           "HubContentArn": "string"
                        },
                        "ManifestEtag": "string",
                        "ManifestS3Uri": "string",
                        "ModelAccessConfig": {
                           "AcceptEula": boolean
                        },
                        "S3DataType": "string",
                        "S3Uri": "string"
                     }
                  },
                  "ModelDataUrl": "string",
                  "ModelInput": {
                     "DataInputConfig": "string"
                  },
                  "NearestModelName": "string",
                  "ProductId": "string"
               }
            ],
            "SupportedContentTypes": [ "string" ],
            "SupportedRealtimeInferenceInstanceTypes": [ "string" ],
            "SupportedResponseMIMETypes": [ "string" ],
            "SupportedTransformInstanceTypes": [ "string" ]
         },
         "ModelApprovalStatus": "string",
         "ModelPackageArn": "string",
         "ModelPackageDescription": "string",
         "ModelPackageGroupName": "string",
         "ModelPackageRegistrationType": "string",
         "ModelPackageStatus": "string",
         "ModelPackageVersion": number
      }
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

BatchDescribeModelPackageErrorMap

A map of the resource and BatchDescribeModelPackageError objects reporting the error associated with describing the model package.

Type: String to BatchDescribeModelPackageError object map

Key Length Constraints: Minimum length of 1. Maximum length of 2048.

Key Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

ModelPackageSummaries

The summaries for the model package versions

Type: String to BatchDescribeModelPackageSummary object map

Key Length Constraints: Minimum length of 1. Maximum length of 2048.

Key Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

BatchRebootClusterNodes

Service: Amazon SageMaker Service

Reboots specific nodes within a SageMaker HyperPod cluster using a soft recovery mechanism. `BatchRebootClusterNodes` performs a graceful reboot of the specified nodes by calling the Amazon Elastic Compute Cloud `RebootInstances` API, which attempts to cleanly shut down the operating system before restarting the instance.

This operation is useful for recovering from transient issues or applying certain configuration changes that require a restart.

> **Note**
>
> - Rebooting a node may cause temporary service interruption for workloads running on that node. Ensure your workloads can handle node restarts or use appropriate scheduling to minimize impact.
> - You can reboot up to 25 nodes in a single request.
> - For SageMaker HyperPod clusters using the Slurm workload manager, ensure rebooting nodes will not disrupt critical cluster operations.

Request Syntax

```
{
   "ClusterName": "string",
   "NodeIds": [ "string" ],
   "NodeLogicalIds": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterName

The name or Amazon Resource Name (ARN) of the SageMaker HyperPod cluster containing the nodes to reboot.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

NodeIds

A list of EC2 instance IDs to reboot using soft recovery. You can specify between 1 and 25 instance IDs.

> **Note**
>
> - Either `NodeIds` or `NodeLogicalIds` must be provided (or both), but at least one is required.
> - Each instance ID must follow the pattern `i-` followed by 17 hexadecimal characters (for example, `i-0123456789abcdef0`).

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: No

NodeLogicalIds

A list of logical node IDs to reboot using soft recovery. You can specify between 1 and 25 logical node IDs.

The `NodeLogicalId` is a unique identifier that persists throughout the node's lifecycle and can be used to track nodes that are still being provisioned and don't yet have an EC2 instance ID assigned.

> **Important**
>
> - This parameter is only supported for clusters using `Continuous` as the `NodeProvisioningMode`. For clusters using the default provisioning mode, use `NodeIds` instead.
> - Either `NodeIds` or `NodeLogicalIds` must be provided (or both), but at least one is required.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Required: No

Response Syntax

```
{
   "Failed": [
      {
         "ErrorCode": "string",
         "Message": "string",
         "NodeId": "string"
      }
   ],
   "FailedNodeLogicalIds": [
      {
         "ErrorCode": "string",
         "Message": "string",
         "NodeLogicalId": "string"
      }
   ],
   "Successful": [ "string" ],
   "SuccessfulNodeLogicalIds": [ "string" ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Failed

A list of errors encountered for EC2 instance IDs that could not be rebooted. Each error includes the instance ID, an error code, and a descriptive message.

Type: Array of BatchRebootClusterNodesError objects

Array Members: Minimum number of 0 items. Maximum number of 25 items.

FailedNodeLogicalIds

A list of errors encountered for logical node IDs that could not be rebooted. Each error includes the logical node ID, an error code, and a descriptive message. This field is only present when `NodeLogicalIds` were provided in the request.

Type: Array of BatchRebootClusterNodeLogicalIdsError objects

Array Members: Minimum number of 0 items. Maximum number of 25 items.

Successful

A list of EC2 instance IDs for which the reboot operation was successfully initiated.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 3000 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

SuccessfulNodeLogicalIds

A list of logical node IDs for which the reboot operation was successfully initiated. This field is only present when `NodeLogicalIds` were provided in the request.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 99 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

BatchReplaceClusterNodes

Service: Amazon SageMaker Service

Replaces specific nodes within a SageMaker HyperPod cluster with new hardware. `BatchReplaceClusterNodes` terminates the specified instances and provisions new replacement instances with the same configuration but fresh hardware. The Amazon Machine Image (AMI) and instance configuration remain the same.

This operation is useful for recovering from hardware failures or persistent issues that cannot be resolved through a reboot.

Important

- **Data Loss Warning:** Replacing nodes destroys all instance volumes, including both root and secondary volumes. All data stored on these volumes will be permanently lost and cannot be recovered.
- To safeguard your work, back up your data to Amazon S3 or an FSx for Lustre file system before invoking the API on a worker node group. This will help prevent any potential data loss from the instance root volume. For more information about backup, see Use the backup script provided by SageMaker HyperPod.
- If you want to invoke this API on an existing cluster, you'll first need to patch the cluster by running the UpdateClusterSoftware API. For more information about patching a cluster, see Update the SageMaker HyperPod platform software of a cluster.
- You can replace up to 25 nodes in a single request.

Request Syntax

```
{
   "ClusterName": "string",
   "NodeIds": [ "string" ],
   "NodeLogicalIds": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterName

The name or Amazon Resource Name (ARN) of the SageMaker HyperPod cluster containing the nodes to replace.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

NodeIds

A list of EC2 instance IDs to replace with new hardware. You can specify between 1 and 25 instance IDs.

> **Important**
>
> Replace operations destroy all instance volumes (root and secondary). Ensure you have backed up any important data before proceeding.

> **Note**
>
> - Either `NodeIds` or `NodeLogicalIds` must be provided (or both), but at least one is required.
> - Each instance ID must follow the pattern `i-` followed by 17 hexadecimal characters (for example, `i-0123456789abcdef0`).
> - For SageMaker HyperPod clusters using the Slurm workload manager, you cannot replace instances that are configured as Slurm controller nodes.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: No

NodeLogicalIds

A list of logical node IDs to replace with new hardware. You can specify between 1 and 25 logical node IDs.

The `NodeLogicalId` is a unique identifier that persists throughout the node's lifecycle and can be used to track nodes that are still being provisioned and don't yet have an EC2 instance ID assigned.

> **Important**
>
> - Replace operations destroy all instance volumes (root and secondary). Ensure you have backed up any important data before proceeding.
> - This parameter is only supported for clusters using `Continuous` as the `NodeProvisioningMode`. For clusters using the default provisioning mode, use `NodeIds` instead.
> - Either `NodeIds` or `NodeLogicalIds` must be provided (or both), but at least one is required.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Required: No

Response Syntax

```
{
   "Failed": [
      {
```

```
         "ErrorCode": "string",
         "Message": "string",
         "NodeId": "string"
      }
   ],
   "FailedNodeLogicalIds": [
      {
         "ErrorCode": "string",
         "Message": "string",
         "NodeLogicalId": "string"
      }
   ],
   "Successful": [ "string" ],
   "SuccessfulNodeLogicalIds": [ "string" ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Failed

A list of errors encountered for EC2 instance IDs that could not be replaced. Each error includes the instance ID, an error code, and a descriptive message.

Type: Array of BatchReplaceClusterNodesError objects

Array Members: Minimum number of 0 items. Maximum number of 25 items.

FailedNodeLogicalIds

A list of errors encountered for logical node IDs that could not be replaced. Each error includes the logical node ID, an error code, and a descriptive message. This field is only present when `NodeLogicalIds` were provided in the request.

Type: Array of BatchReplaceClusterNodeLogicalIdsError objects

Array Members: Minimum number of 0 items. Maximum number of 25 items.

Successful

A list of EC2 instance IDs for which the replacement operation was successfully initiated.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 3000 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

SuccessfulNodeLogicalIds

A list of logical node IDs for which the replacement operation was successfully initiated. This field is only present when `NodeLogicalIds` were provided in the request.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 99 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateAction

Service: Amazon SageMaker Service

Creates an *action*. An action is a lineage tracking entity that represents an action or activity. For example, a model deployment or an HPO job. Generally, an action involves at least one input or output artifact. For more information, see Amazon SageMaker ML Lineage Tracking.

Request Syntax

```
{
   "ActionName": "string",
   "ActionType": "string",
   "Description": "string",
   "MetadataProperties": {
      "CommitId": "string",
      "GeneratedBy": "string",
      "ProjectId": "string",
      "Repository": "string"
   },
   "Properties": {
      "string" : "string"
   },
   "Source": {
      "SourceId": "string",
      "SourceType": "string",
      "SourceUri": "string"
   },
   "Status": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ActionName

The name of the action. Must be unique to your account in an AWS Region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

ActionType

The action type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

Description

The description of the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Required: No

Properties

A list of properties to add to the action.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 30 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2500.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Required: No

Source

The source type, ID, and URI.

Type: ActionSource object

Required: Yes

Status

The status of the action.

Type: String

Valid Values: `Unknown | InProgress | Completed | Failed | Stopping | Stopped`

Required: No

Tags

A list of tags to apply to the action.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "ActionArn": "string"
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ActionArn

The Amazon Resource Name (ARN) of the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:action/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateAIBenchmarkJob

Service: Amazon SageMaker Service

Creates a benchmark job that runs performance benchmarks against inference infrastructure using a predefined AI workload configuration. The benchmark job measures metrics such as latency, throughput, and cost for your generative AI inference endpoints.

Request Syntax

```
{
   "AIBenchmarkJobName": "string",
   "AIWorkloadConfigIdentifier": "string",
   "BenchmarkTarget": { ... },
   "NetworkConfig": {
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
   },
   "OutputConfig": {
      "MlflowConfig": {
         "MlflowExperimentName": "string",
         "MlflowResourceArn": "string",
         "MlflowRunName": "string"
      },
      "S3OutputLocation": "string"
   },
   "RoleArn": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AIBenchmarkJobName

The name of the AI benchmark job. The name must be unique within your AWS account in the current AWS Region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AIWorkloadConfigIdentifier

The name or Amazon Resource Name (ARN) of the AI workload configuration to use for this benchmark job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*/)?([a-zA-Z0-9]([a-zA-Z0-9\-]){0,62})(?<!-)`

Required: Yes

BenchmarkTarget

The target endpoint to benchmark. Specify a SageMaker endpoint by providing its name or Amazon Resource Name (ARN).

Type: AIBenchmarkTarget object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

NetworkConfig

The network configuration for the benchmark job, including VPC settings.

Type: AIBenchmarkNetworkConfig object

Required: No

OutputConfig

The output configuration for the benchmark job, including the Amazon S3 location where benchmark results are stored.

Type: AIBenchmarkOutputConfig object

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of an IAM role that enables Amazon SageMaker AI to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Tags

The metadata that you apply to AWS resources to help you categorize and organize them. Each tag consists of a key and a value, both of which you define.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "AIBenchmarkJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIBenchmarkJobArn

The Amazon Resource Name (ARN) of the created benchmark job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-benchmark-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateAIRecommendationJob

Service: Amazon SageMaker Service

Creates a recommendation job that generates intelligent optimization recommendations for generative AI inference deployments. The job analyzes your model, workload configuration, and performance targets to recommend optimal instance types, model optimization techniques (such as quantization and speculative decoding), and deployment configurations.

Request Syntax

```
{
   "AIRecommendationJobName": "string",
   "AIWorkloadConfigIdentifier": "string",
   "ComputeSpec": {
      "CapacityReservationConfig": {
         "CapacityReservationPreference": "string",
         "MlReservationArns": [ "string" ]
      },
      "InstanceTypes": [ "string" ]
   },
   "InferenceSpecification": {
      "Framework": "string"
   },
   "ModelSource": { ... },
   "OptimizeModel": boolean,
   "OutputConfig": {
      "MlflowConfig": {
         "MlflowExperimentName": "string",
         "MlflowResourceArn": "string",
         "MlflowRunName": "string"
      },
      "ModelPackageGroupIdentifier": "string",
      "S3OutputLocation": "string"
   },
   "PerformanceTarget": {
      "Constraints": [
         {
            "Metric": "string"
         }
      ]
   },
   "RoleArn": "string",
```

```
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AIRecommendationJobName

The name of the AI recommendation job. The name must be unique within your AWS account in the current AWS Region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AIWorkloadConfigIdentifier

The name or Amazon Resource Name (ARN) of the AI workload configuration to use for this recommendation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*/)?([a-zA-Z0-9]([a-zA-Z0-9\-]){0,62})(?<!-)`

Required: Yes

ComputeSpec

The compute resource specification for the recommendation job. You can specify up to 3 instance types to consider, and optionally provide capacity reservation configuration.

Type: AIRecommendationComputeSpec object

Required: No

InferenceSpecification

The inference framework configuration. Specify the framework (such as LMI or vLLM) for the recommendation job.

Type: AIRecommendationInferenceSpecification object

Required: No

ModelSource

The source of the model to optimize. Specify the Amazon S3 location of the model artifacts.

Type: AIModelSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

OptimizeModel

Whether to allow model optimization techniques such as quantization, speculative decoding, and kernel tuning. The default is `true`.

Type: Boolean

Required: No

OutputConfig

The output configuration for the recommendation job, including the Amazon S3 location for results and an optional model package group where the optimized model is registered.

Type: AIRecommendationOutputConfig object

Required: Yes

PerformanceTarget

The performance targets for the recommendation job. Specify constraints on metrics such as time to first token (`ttft-ms`), `throughput`, or `cost`.

Type: AIRecommendationPerformanceTarget object

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of an IAM role that enables Amazon SageMaker AI to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Tags

The metadata that you apply to AWS resources to help you categorize and organize them.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "AIRecommendationJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIRecommendationJobArn

The Amazon Resource Name (ARN) of the created recommendation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-recommendation-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

CreateAIWorkloadConfig

Service: Amazon SageMaker Service

Creates a reusable AI workload configuration that defines datasets, data sources, and benchmark tool settings for consistent performance testing of generative AI inference deployments on Amazon SageMaker AI.

Request Syntax

```
{
   "AIWorkloadConfigName": "string",
   "AIWorkloadConfigs": {
      "WorkloadSpec": { ... }
   },
   "DatasetConfig": { ... },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AIWorkloadConfigName

The name of the AI workload configuration. The name must be unique within your AWS account in the current AWS Region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AIWorkloadConfigs

The benchmark tool configuration and workload specification. Provide the specification as an inline YAML or JSON string.

Type: AIWorkloadConfigs object

Required: No

DatasetConfig

The dataset configuration for the workload. Specify input data channels with their data sources for benchmark workloads.

Type: AIDatasetConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

Tags

The metadata that you apply to AWS resources to help you categorize and organize them. Each tag consists of a key and a value, both of which you define. For more information, see Tagging AWS Resources in the AWS General Reference.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "AIWorkloadConfigArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIWorkloadConfigArn

The Amazon Resource Name (ARN) of the created AI workload configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-workload-config/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateAlgorithm

Service: Amazon SageMaker Service

Create a machine learning algorithm that you can use in SageMaker and list in the AWS Marketplace.

Request Syntax

```
{
   "AlgorithmDescription": "string",
   "AlgorithmName": "string",
   "CertifyForMarketplace": boolean,
   "InferenceSpecification": {
      "Containers": [
         {
            "AdditionalModelDataSources": [
               {
                  "ChannelName": "string",
                  "S3DataSource": {
                     "CompressionType": "string",
                     "ETag": "string",
                     "HubAccessConfig": {
                        "HubContentArn": "string"
                     },
                     "ManifestEtag": "string",
                     "ManifestS3Uri": "string",
                     "ModelAccessConfig": {
                        "AcceptEula": boolean
                     },
                     "S3DataType": "string",
                     "S3Uri": "string"
                  }
               }
            ],
            "AdditionalS3DataSource": {
               "CompressionType": "string",
               "ETag": "string",
               "S3DataType": "string",
               "S3Uri": "string"
            },
            "BaseModel": {
               "HubContentName": "string",
               "HubContentVersion": "string",
```

```
            "RecipeName": "string"
         },
         "ContainerHostname": "string",
         "Environment": {
            "string" : "string"
         },
         "Framework": "string",
         "FrameworkVersion": "string",
         "Image": "string",
         "ImageDigest": "string",
         "IsCheckpoint": boolean,
         "ModelDataETag": "string",
         "ModelDataSource": {
            "S3DataSource": {
               "CompressionType": "string",
               "ETag": "string",
               "HubAccessConfig": {
                  "HubContentArn": "string"
               },
               "ManifestEtag": "string",
               "ManifestS3Uri": "string",
               "ModelAccessConfig": {
                  "AcceptEula": boolean
               },
               "S3DataType": "string",
               "S3Uri": "string"
            }
         },
         "ModelDataUrl": "string",
         "ModelInput": {
            "DataInputConfig": "string"
         },
         "NearestModelName": "string",
         "ProductId": "string"
      }
   ],
   "SupportedContentTypes": [ "string" ],
   "SupportedRealtimeInferenceInstanceTypes": [ "string" ],
   "SupportedResponseMIMETypes": [ "string" ],
   "SupportedTransformInstanceTypes": [ "string" ]
},
"Tags": [
   {
      "Key": "string",
```

```
            "Value": "string"
         }
      ],
      "TrainingSpecification": {
         "AdditionalS3DataSource": {
            "CompressionType": "string",
            "ETag": "string",
            "S3DataType": "string",
            "S3Uri": "string"
         },
         "MetricDefinitions": [
            {
               "Name": "string",
               "Regex": "string"
            }
         ],
         "SupportedHyperParameters": [
            {
               "DefaultValue": "string",
               "Description": "string",
               "IsRequired": boolean,
               "IsTunable": boolean,
               "Name": "string",
               "Range": {
                  "CategoricalParameterRangeSpecification": {
                     "Values": [ "string" ]
                  },
                  "ContinuousParameterRangeSpecification": {
                     "MaxValue": "string",
                     "MinValue": "string"
                  },
                  "IntegerParameterRangeSpecification": {
                     "MaxValue": "string",
                     "MinValue": "string"
                  }
               },
               "Type": "string"
            }
         ],
         "SupportedTrainingInstanceTypes": [ "string" ],
         "SupportedTuningJobObjectiveMetrics": [
            {
               "MetricName": "string",
               "Type": "string"
```

```
            }
         ],
         "SupportsDistributedTraining": boolean,
         "TrainingChannels": [
            {
               "Description": "string",
               "IsRequired": boolean,
               "Name": "string",
               "SupportedCompressionTypes": [ "string" ],
               "SupportedContentTypes": [ "string" ],
               "SupportedInputModes": [ "string" ]
            }
         ],
         "TrainingImage": "string",
         "TrainingImageDigest": "string"
      },
      "ValidationSpecification": {
         "ValidationProfiles": [
            {
               "ProfileName": "string",
               "TrainingJobDefinition": {
                  "HyperParameters": {
                     "string" : "string"
                  },
                  "InputDataConfig": [
                     {
                        "ChannelName": "string",
                        "CompressionType": "string",
                        "ContentType": "string",
                        "DataSource": {
                           "DatasetSource": {
                              "DatasetArn": "string"
                           },
                           "FileSystemDataSource": {
                              "DirectoryPath": "string",
                              "FileSystemAccessMode": "string",
                              "FileSystemId": "string",
                              "FileSystemType": "string"
                           },
                           "S3DataSource": {
                              "AttributeNames": [ "string" ],
                              "HubAccessConfig": {
                                 "HubContentArn": "string"
                              },
```

```
                        "InstanceGroupNames": [ "string" ],
                        "ModelAccessConfig": {
                           "AcceptEula": boolean
                        },
                        "S3DataDistributionType": "string",
                        "S3DataType": "string",
                        "S3Uri": "string"
                     }
                  },
                  "InputMode": "string",
                  "RecordWrapperType": "string",
                  "ShuffleConfig": {
                     "Seed": number
                  }
               }
            ],
            "OutputDataConfig": {
               "CompressionType": "string",
               "KmsKeyId": "string",
               "S3OutputPath": "string"
            },
            "ResourceConfig": {
               "InstanceCount": number,
               "InstanceGroups": [
                  {
                     "InstanceCount": number,
                     "InstanceGroupName": "string",
                     "InstanceType": "string"
                  }
               ],
               "InstancePlacementConfig": {
                  "EnableMultipleJobs": boolean,
                  "PlacementSpecifications": [
                     {
                        "InstanceCount": number,
                        "UltraServerId": "string"
                     }
                  ]
               },
               "InstanceType": "string",
               "KeepAlivePeriodInSeconds": number,
               "TrainingPlanArn": "string",
               "VolumeKmsKeyId": "string",
               "VolumeSizeInGB": number
```

```
            },
            "StoppingCondition": {
               "MaxPendingTimeInSeconds": number,
               "MaxRuntimeInSeconds": number,
               "MaxWaitTimeInSeconds": number
            },
            "TrainingInputMode": "string"
         },
         "TransformJobDefinition": {
            "BatchStrategy": "string",
            "Environment": {
               "string" : "string"
            },
            "MaxConcurrentTransforms": number,
            "MaxPayloadInMB": number,
            "TransformInput": {
               "CompressionType": "string",
               "ContentType": "string",
               "DataSource": {
                  "S3DataSource": {
                     "S3DataType": "string",
                     "S3Uri": "string"
                  }
               },
               "SplitType": "string"
            },
            "TransformOutput": {
               "Accept": "string",
               "AssembleWith": "string",
               "KmsKeyId": "string",
               "S3OutputPath": "string"
            },
            "TransformResources": {
               "InstanceCount": number,
               "InstanceType": "string",
               "TransformAmiVersion": "string",
               "VolumeKmsKeyId": "string"
            }
         }
      }
   ],
   "ValidationRole": "string"
}
```

```
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AlgorithmDescription

A description of the algorithm.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

AlgorithmName

The name of the algorithm.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

CertifyForMarketplace

Whether to certify the algorithm so that it can be listed in AWS Marketplace.

Type: Boolean

Required: No

InferenceSpecification

Specifies details about inference jobs that the algorithm runs, including the following:

- The Amazon ECR paths of containers that contain the inference code and model artifacts.
- The instance types that the algorithm supports for transform jobs and real-time endpoints used for inference.

- The input and output content formats that the algorithm supports for inference.

Type: InferenceSpecification object

Required: No

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TrainingSpecification

Specifies details about training jobs run by this algorithm, including the following:

- The Amazon ECR path of the container and the version digest of the algorithm.
- The hyperparameters that the algorithm supports.
- The instance types that the algorithm supports for training.
- Whether the algorithm supports distributed training.
- The metrics that the algorithm emits to Amazon CloudWatch.
- Which metrics that the algorithm emits can be used as the objective metric for hyperparameter tuning jobs.
- The input channels that the algorithm supports for training data. For example, an algorithm might support `train`, `validation`, and `test` channels.

Type: TrainingSpecification object

Required: Yes

ValidationSpecification

Specifies configurations for one or more training jobs and that SageMaker runs to test the algorithm's training code and, optionally, one or more batch transform jobs that SageMaker runs to test the algorithm's inference code.

Type: AlgorithmValidationSpecification object

Required: No

Response Syntax

```
{
   "AlgorithmArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AlgorithmArn

The Amazon Resource Name (ARN) of the new algorithm.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:algorithm/[\S]{1,2048}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateApp

Service: Amazon SageMaker Service

Creates a running app for the specified UserProfile. This operation is automatically invoked by Amazon SageMaker AI upon access to the associated Domain, and when new kernel configurations are selected by the user. A user may have multiple Apps active simultaneously.

Request Syntax

```
{
   "AppName": "string",
   "AppType": "string",
   "DomainId": "string",
   "RecoveryMode": boolean,
   "ResourceSpec": {
      "InstanceType": "string",
      "LifecycleConfigArn": "string",
      "SageMakerImageArn": "string",
      "SageMakerImageVersionAlias": "string",
      "SageMakerImageVersionArn": "string",
      "TrainingPlanArn": "string"
   },
   "SpaceName": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "UserProfileName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AppName

The name of the app.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AppType

The type of app.

Type: String

Valid Values: `JupyterServer | KernelGateway | DetailedProfiler | TensorBoard | CodeEditor | JupyterLab | RStudioServerPro | RSessionGateway | Canvas`

Required: Yes

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

RecoveryMode

Indicates whether the application is launched in recovery mode.

Type: Boolean

Required: No

ResourceSpec

The instance type and the Amazon Resource Name (ARN) of the SageMaker AI image created on the instance.

> **Note**
>
> The value of `InstanceType` passed as part of the `ResourceSpec` in the `CreateApp` call overrides the value passed as part of the `ResourceSpec` configured for the user profile or the domain. If `InstanceType` is not specified in any of those three

`ResourceSpec` values for a `KernelGateway` app, the `CreateApp` call fails with a request validation error.

Type: ResourceSpec object

Required: No

SpaceName

The name of the space. If this value is not set, then `UserProfileName` must be set.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Tags

Each tag consists of a key and an optional value. Tag keys must be unique per resource.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

UserProfileName

The user profile name. If this value is not set, then `SpaceName` must be set.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Response Syntax

```
{
```

```
   "AppArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AppArn

The Amazon Resource Name (ARN) of the app.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:app/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateAppImageConfig

Service: Amazon SageMaker Service

Creates a configuration for running a SageMaker AI image as a KernelGateway app. The configuration specifies the Amazon Elastic File System storage volume on the image, and a list of the kernels in the image.

Request Syntax

```
{
   "AppImageConfigName": "string",
   "CodeEditorAppImageConfig": {
      "ContainerConfig": {
         "ContainerArguments": [ "string" ],
         "ContainerEntrypoint": [ "string" ],
         "ContainerEnvironmentVariables": {
            "string" : "string"
         }
      },
      "FileSystemConfig": {
         "DefaultGid": number,
         "DefaultUid": number,
         "MountPath": "string"
      }
   },
   "JupyterLabAppImageConfig": {
      "ContainerConfig": {
         "ContainerArguments": [ "string" ],
         "ContainerEntrypoint": [ "string" ],
         "ContainerEnvironmentVariables": {
            "string" : "string"
         }
      },
      "FileSystemConfig": {
         "DefaultGid": number,
         "DefaultUid": number,
         "MountPath": "string"
      }
   },
   "KernelGatewayImageConfig": {
      "FileSystemConfig": {
         "DefaultGid": number,
         "DefaultUid": number,
```

```
            "MountPath": "string"
         },
         "KernelSpecs": [
            {
               "DisplayName": "string",
               "Name": "string"
            }
         ]
      },
      "Tags": [
         {
            "Key": "string",
            "Value": "string"
         }
      ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AppImageConfigName

The name of the AppImageConfig. Must be unique to your account.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

CodeEditorAppImageConfig

The `CodeEditorAppImageConfig`. You can only specify one image kernel in the AppImageConfig API. This kernel is shown to users before the image starts. After the image runs, all kernels are visible in Code Editor.

Type: CodeEditorAppImageConfig object

Required: No

JupyterLabAppImageConfig

The `JupyterLabAppImageConfig`. You can only specify one image kernel in the `AppImageConfig` API. This kernel is shown to users before the image starts. After the image runs, all kernels are visible in JupyterLab.

Type: JupyterLabAppImageConfig object

Required: No

KernelGatewayImageConfig

The KernelGatewayImageConfig. You can only specify one image kernel in the AppImageConfig API. This kernel will be shown to users before the image starts. Once the image runs, all kernels are visible in JupyterLab.

Type: KernelGatewayImageConfig object

Required: No

Tags

A list of tags to apply to the AppImageConfig.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "AppImageConfigArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AppImageConfigArn

The ARN of the AppImageConfig.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:app-image-config/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateArtifact

Service: Amazon SageMaker Service

Creates an *artifact*. An artifact is a lineage tracking entity that represents a URI addressable object or data. Some examples are the S3 URI of a dataset and the ECR registry path of an image. For more information, see Amazon SageMaker ML Lineage Tracking.

Request Syntax

```
{
   "ArtifactName": "string",
   "ArtifactType": "string",
   "MetadataProperties": {
      "CommitId": "string",
      "GeneratedBy": "string",
      "ProjectId": "string",
      "Repository": "string"
   },
   "Properties": {
      "string" : "string"
   },
   "Source": {
      "SourceTypes": [
         {
            "SourceIdType": "string",
            "Value": "string"
         }
      ],
      "SourceUri": "string"
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ArtifactName

The name of the artifact. Must be unique to your account in an AWS Region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

ArtifactType

The artifact type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Required: No

Properties

A list of properties to add to the artifact.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 30 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2500.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 4096.

Value Pattern: `.*`

Required: No

Source

The ID, ID type, and URI of the source.

Type: ArtifactSource object

Required: Yes

Tags

A list of tags to apply to the artifact.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "ArtifactArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ArtifactArn

The Amazon Resource Name (ARN) of the artifact.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:artifact/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateAutoMLJob

Service: Amazon SageMaker Service

Creates an Autopilot job also referred to as Autopilot experiment or AutoML job.

An AutoML job in SageMaker AI is a fully automated process that allows you to build machine learning models with minimal effort and machine learning expertise. When initiating an AutoML job, you provide your data and optionally specify parameters tailored to your use case. SageMaker AI then automates the entire model development lifecycle, including data preprocessing, model training, tuning, and evaluation. AutoML jobs are designed to simplify and accelerate the model building process by automating various tasks and exploring different combinations of machine learning algorithms, data preprocessing techniques, and hyperparameter values. The output of an AutoML job comprises one or more trained models ready for deployment and inference. Additionally, SageMaker AI AutoML jobs generate a candidate model leaderboard, allowing you to select the best-performing model for deployment.

For more information about AutoML jobs, see https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-automate-model-development.html in the SageMaker AI developer guide.

> **Note**
>
> We recommend using the new versions CreateAutoMLJobV2 and DescribeAutoMLJobV2, which offer backward compatibility.
> `CreateAutoMLJobV2` can manage tabular problem types identical to those of its previous version `CreateAutoMLJob`, as well as time-series forecasting, non-tabular problem types such as image or text classification, and text generation (LLMs fine-tuning).
> Find guidelines about how to migrate a `CreateAutoMLJob` to `CreateAutoMLJobV2` in Migrate a CreateAutoMLJob to CreateAutoMLJobV2.

You can find the best-performing model after you run an AutoML job by calling DescribeAutoMLJobV2 (recommended) or DescribeAutoMLJob.

Request Syntax

```
{
   "AutoMLJobConfig": {
      "CandidateGenerationConfig": {
         "AlgorithmsConfig": [
```

```
            {
               "AutoMLAlgorithms": [ "string" ]
            }
         ],
         "FeatureSpecificationS3Uri": "string"
      },
      "CompletionCriteria": {
         "MaxAutoMLJobRuntimeInSeconds": number,
         "MaxCandidates": number,
         "MaxRuntimePerTrainingJobInSeconds": number
      },
      "DataSplitConfig": {
         "ValidationFraction": number
      },
      "Mode": "string",
      "SecurityConfig": {
         "EnableInterContainerTrafficEncryption": boolean,
         "VolumeKmsKeyId": "string",
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      }
   },
   "AutoMLJobName": "string",
   "AutoMLJobObjective": {
      "MetricName": "string"
   },
   "GenerateCandidateDefinitionsOnly": boolean,
   "InputDataConfig": [
      {
         "ChannelType": "string",
         "CompressionType": "string",
         "ContentType": "string",
         "DataSource": {
            "S3DataSource": {
               "S3DataType": "string",
               "S3Uri": "string"
            }
         },
         "SampleWeightAttributeName": "string",
         "TargetAttributeName": "string"
      }
   ],
```

```
   "ModelDeployConfig": {
      "AutoGenerateEndpointName": boolean,
      "EndpointName": "string"
   },
   "OutputDataConfig": {
      "KmsKeyId": "string",
      "S3OutputPath": "string"
   },
   "ProblemType": "string",
   "RoleArn": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AutoMLJobConfig

A collection of settings used to configure an AutoML job.

Type: AutoMLJobConfig object

Required: No

AutoMLJobName

Identifies an Autopilot job. The name must be unique to your account and is case insensitive.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

AutoMLJobObjective

Specifies a metric to minimize or maximize as the objective of a job. If not specified, the default objective metric depends on the problem type. See AutoMLJobObjective for the default values.

Type: AutoMLJobObjective object

Required: No

GenerateCandidateDefinitionsOnly

Generates possible candidates without training the models. A candidate is a combination of data preprocessors, algorithms, and algorithm parameter settings.

Type: Boolean

Required: No

InputDataConfig

An array of channel objects that describes the input data and its location. Each channel is a named input source. Similar to `InputDataConfig` supported by HyperParameterTrainingJobDefinition. Format(s) supported: CSV, Parquet. A minimum of 500 rows is required for the training dataset. There is not a minimum number of rows required for the validation dataset.

Type: Array of AutoMLChannel objects

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Required: Yes

ModelDeployConfig

Specifies how to generate the endpoint name for an automatic one-click Autopilot model deployment.

Type: ModelDeployConfig object

Required: No

OutputDataConfig

Provides information about encryption and the Amazon S3 output path needed to store artifacts from an AutoML job. Format(s) supported: CSV.

Type: AutoMLOutputDataConfig object

Required: Yes

ProblemType

Defines the type of supervised learning problem available for the candidates. For more information, see SageMaker Autopilot problem types.

Type: String

Valid Values: `BinaryClassification | MulticlassClassification | Regression`

Required: No

RoleArn

The ARN of the role that is used to access the data.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWSResources. Tag keys must be unique per resource.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "AutoMLJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AutoMLJobArn

The unique ARN assigned to the AutoML job when it is created.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateAutoMLJobV2

Service: Amazon SageMaker Service

Creates an Autopilot job also referred to as Autopilot experiment or AutoML job V2.

An AutoML job in SageMaker AI is a fully automated process that allows you to build machine learning models with minimal effort and machine learning expertise. When initiating an AutoML job, you provide your data and optionally specify parameters tailored to your use case. SageMaker AI then automates the entire model development lifecycle, including data preprocessing, model training, tuning, and evaluation. AutoML jobs are designed to simplify and accelerate the model building process by automating various tasks and exploring different combinations of machine learning algorithms, data preprocessing techniques, and hyperparameter values. The output of an AutoML job comprises one or more trained models ready for deployment and inference. Additionally, SageMaker AI AutoML jobs generate a candidate model leaderboard, allowing you to select the best-performing model for deployment.

For more information about AutoML jobs, see https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-automate-model-development.html in the SageMaker AI developer guide.

AutoML jobs V2 support various problem types such as regression, binary, and multiclass classification with tabular data, text and image classification, time-series forecasting, and fine-tuning of large language models (LLMs) for text generation.

> **Note**
>
> CreateAutoMLJobV2 and DescribeAutoMLJobV2 are new versions of CreateAutoMLJob and DescribeAutoMLJob which offer backward compatibility.
> `CreateAutoMLJobV2` can manage tabular problem types identical to those of its previous version `CreateAutoMLJob`, as well as time-series forecasting, non-tabular problem types such as image or text classification, and text generation (LLMs fine-tuning).
> Find guidelines about how to migrate a `CreateAutoMLJob` to `CreateAutoMLJobV2` in Migrate a CreateAutoMLJob to CreateAutoMLJobV2.

For the list of available problem types supported by `CreateAutoMLJobV2`, see AutoMLProblemTypeConfig.

You can find the best-performing model after you run an AutoML job V2 by calling DescribeAutoMLJobV2.

Request Syntax

```
{
   "AutoMLComputeConfig": {
      "EmrServerlessComputeConfig": {
         "ExecutionRoleARN": "string"
      }
   },
   "AutoMLJobInputDataConfig": [
      {
         "ChannelType": "string",
         "CompressionType": "string",
         "ContentType": "string",
         "DataSource": {
            "S3DataSource": {
               "S3DataType": "string",
               "S3Uri": "string"
            }
         }
      }
   ],
   "AutoMLJobName": "string",
   "AutoMLJobObjective": {
      "MetricName": "string"
   },
   "AutoMLProblemTypeConfig": { ... },
   "DataSplitConfig": {
      "ValidationFraction": number
   },
   "ModelDeployConfig": {
      "AutoGenerateEndpointName": boolean,
      "EndpointName": "string"
   },
   "OutputDataConfig": {
      "KmsKeyId": "string",
      "S3OutputPath": "string"
   },
   "RoleArn": "string",
   "SecurityConfig": {
      "EnableInterContainerTrafficEncryption": boolean,
      "VolumeKmsKeyId": "string",
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
```

```
            "Subnets": [ "string" ]
        }
    },
    "Tags": [
        {
            "Key": "string",
            "Value": "string"
        }
    ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AutoMLComputeConfig

Specifies the compute configuration for the AutoML job V2.

Type: AutoMLComputeConfig object

Required: No

AutoMLJobInputDataConfig

An array of channel objects describing the input data and their location. Each channel is a named input source. Similar to the InputDataConfig attribute in the `CreateAutoMLJob` input parameters. The supported formats depend on the problem type:

- For tabular problem types: `S3Prefix`, `ManifestFile`.
- For image classification: `S3Prefix`, `ManifestFile`, `AugmentedManifestFile`.
- For text classification: `S3Prefix`.
- For time-series forecasting: `S3Prefix`.
- For text generation (LLMs fine-tuning): `S3Prefix`.

Type: Array of AutoMLJobChannel objects

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Required: Yes

AutoMLJobName

Identifies an Autopilot job. The name must be unique to your account and is case insensitive.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

AutoMLJobObjective

Specifies a metric to minimize or maximize as the objective of a job. If not specified, the default objective metric depends on the problem type. For the list of default values per problem type, see AutoMLJobObjective.

> **Note**
>
> - For tabular problem types: You must either provide both the `AutoMLJobObjective` and indicate the type of supervised learning problem in `AutoMLProblemTypeConfig` (`TabularJobConfig.ProblemType`), or none at all.
> - For text generation problem types (LLMs fine-tuning): Fine-tuning language models in Autopilot does not require setting the `AutoMLJobObjective` field. Autopilot fine-tunes LLMs without requiring multiple candidates to be trained and evaluated. Instead, using your dataset, Autopilot directly fine-tunes your target model to enhance a default objective metric, the cross-entropy loss. After fine-tuning a language model, you can evaluate the quality of its generated text using different metrics. For a list of the available metrics, see Metrics for fine-tuning LLMs in Autopilot.

Type: AutoMLJobObjective object

Required: No

AutoMLProblemTypeConfig

Defines the configuration settings of one of the supported problem types.

Type: AutoMLProblemTypeConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

DataSplitConfig

This structure specifies how to split the data into train and validation datasets.

The validation and training datasets must contain the same headers. For jobs created by calling `CreateAutoMLJob`, the validation dataset must be less than 2 GB in size.

> **Note**
>
> This attribute must not be set for the time-series forecasting problem type, as Autopilot automatically splits the input dataset into training and validation sets.

Type: AutoMLDataSplitConfig object

Required: No

ModelDeployConfig

Specifies how to generate the endpoint name for an automatic one-click Autopilot model deployment.

Type: ModelDeployConfig object

Required: No

OutputDataConfig

Provides information about encryption and the Amazon S3 output path needed to store artifacts from an AutoML job.

Type: AutoMLOutputDataConfig object

Required: Yes

RoleArn

The ARN of the role that is used to access the data.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

SecurityConfig

The security configuration for traffic encryption or Amazon VPC settings.

Type: AutoMLSecurityConfig object

Required: No

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, such as by purpose, owner, or environment. For more information, see Tagging AWSResources. Tag keys must be unique per resource.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "AutoMLJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AutoMLJobArn

The unique ARN assigned to the AutoMLJob when it is created.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateCluster

Service: Amazon SageMaker Service

Creates an Amazon SageMaker HyperPod cluster. SageMaker HyperPod is a capability of SageMaker for creating and managing persistent clusters for developing large machine learning models, such as large language models (LLMs) and diffusion models. To learn more, see Amazon SageMaker HyperPod in the *Amazon SageMaker Developer Guide*.

Request Syntax

```
{
   "AutoScaling": {
      "AutoScalerType": "string",
      "Mode": "string"
   },
   "ClusterName": "string",
   "ClusterRole": "string",
   "InstanceGroups": [
      {
         "CapacityRequirements": {
            "OnDemand": {
            },
            "Spot": {
            }
         },
         "ExecutionRole": "string",
         "ImageId": "string",
         "InstanceCount": number,
         "InstanceGroupName": "string",
         "InstanceRequirements": {
            "InstanceTypes": [ "string" ]
         },
         "InstanceStorageConfigs": [
            { ... }
         ],
         "InstanceType": "string",
         "KubernetesConfig": {
            "Labels": {
               "string" : "string"
            },
            "Taints": [
               {
```

```
                "Effect": "string",
                "Key": "string",
                "Value": "string"
             }
          ]
       },
       "LifeCycleConfig": {
          "OnCreate": "string",
          "OnInitComplete": "string",
          "SourceS3Uri": "string"
       },
       "MinInstanceCount": number,
       "NetworkInterface": {
          "InterfaceType": "string"
       },
       "OnStartDeepHealthChecks": [ "string" ],
       "OverrideVpcConfig": {
          "SecurityGroupIds": [ "string" ],
          "Subnets": [ "string" ]
       },
       "ScheduledUpdateConfig": {
          "DeploymentConfig": {
             "AutoRollbackConfiguration": [
                {
                   "AlarmName": "string"
                }
             ],
             "RollingUpdatePolicy": {
                "MaximumBatchSize": {
                   "Type": "string",
                   "Value": number
                },
                "RollbackMaximumBatchSize": {
                   "Type": "string",
                   "Value": number
                }
             },
             "WaitIntervalInSeconds": number
          },
          "ScheduleExpression": "string"
       },
       "SlurmConfig": {
          "NodeType": "string",
          "PartitionNames": [ "string" ]
```

```
        },
        "ThreadsPerCore": number,
        "TrainingPlanArn": "string"
      }
   ],
   "NodeProvisioningMode": "string",
   "NodeRecovery": "string",
   "Orchestrator": {
      "Eks": {
         "ClusterArn": "string"
      },
      "Slurm": {
         "SlurmConfigStrategy": "string"
      }
   },
   "RestrictedInstanceGroups": [
      {
         "EnvironmentConfig": {
            "FSxLustreConfig": {
               "PerUnitStorageThroughput": number,
               "SizeInGiB": number
            }
         },
         "ExecutionRole": "string",
         "InstanceCount": number,
         "InstanceGroupName": "string",
         "InstanceStorageConfigs": [
            { ... }
         ],
         "InstanceType": "string",
         "OnStartDeepHealthChecks": [ "string" ],
         "OverrideVpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         },
         "ScheduledUpdateConfig": {
            "DeploymentConfig": {
               "AutoRollbackConfiguration": [
                  {
                     "AlarmName": "string"
                  }
               ],
               "RollingUpdatePolicy": {
                  "MaximumBatchSize": {
```

```
                        "Type": "string",
                        "Value": number
                     },
                     "RollbackMaximumBatchSize": {
                        "Type": "string",
                        "Value": number
                     }
                  },
                  "WaitIntervalInSeconds": number
               },
               "ScheduleExpression": "string"
            },
            "ThreadsPerCore": number,
            "TrainingPlanArn": "string"
         }
      ],
      "RestrictedInstanceGroupsConfig": {
         "SharedEnvironmentConfig": {
            "FSxLustreConfig": {
               "PerUnitStorageThroughput": number,
               "SizeInGiB": number
            },
            "FSxLustreDeletionPolicy": "string"
         }
      },
      "Tags": [
         {
            "Key": "string",
            "Value": "string"
         }
      ],
      "TieredStorageConfig": {
         "InstanceMemoryAllocationPercentage": number,
         "Mode": "string"
      },
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AutoScaling

The autoscaling configuration for the cluster. Enables automatic scaling of cluster nodes based on workload demand using a Karpenter-based system.

Type: ClusterAutoScalingConfig object

Required: No

ClusterName

The name for the new SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

ClusterRole

The Amazon Resource Name (ARN) of the IAM role that HyperPod assumes to perform cluster autoscaling operations. This role must have permissions for `sagemaker:BatchAddClusterNodes` and `sagemaker:BatchDeleteClusterNodes`. This is only required when autoscaling is enabled and when HyperPod is performing autoscaling operations.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

InstanceGroups

The instance groups to be created in the SageMaker HyperPod cluster.

Type: Array of ClusterInstanceGroupSpecification objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

NodeProvisioningMode

The mode for provisioning nodes in the cluster. You can specify the following modes:

- **Continuous**: Scaling behavior that enables 1) concurrent operation execution within instance groups, 2) continuous retry mechanisms for failed operations, 3) enhanced customer visibility into cluster events through detailed event streams, 4) partial provisioning capabilities. Your clusters and instance groups remain `InService` while scaling. This mode is only supported for EKS orchestrated clusters.

Type: String

Valid Values: `Continuous`

Required: No

NodeRecovery

The node recovery mode for the SageMaker HyperPod cluster. When set to `Automatic`, SageMaker HyperPod will automatically reboot or replace faulty nodes when issues are detected. When set to `None`, cluster administrators will need to manually manage any faulty cluster instances.

Type: String

Valid Values: `Automatic | None`

Required: No

Orchestrator

The type of orchestrator to use for the SageMaker HyperPod cluster. Currently, supported values are `"Eks"` and `"Slurm"`, which is to use an Amazon Elastic Kubernetes Service or Slurm cluster as the orchestrator.

> **Note**
>
> If you specify the `Orchestrator` field, you must provide exactly one orchestrator configuration: either `Eks` or `Slurm`. Specifying both or providing an empty configuration returns a validation error.

Type: ClusterOrchestrator object

Required: No

RestrictedInstanceGroups

The specialized instance groups for training models like Amazon Nova to be created in the SageMaker HyperPod cluster.

Type: Array of ClusterRestrictedInstanceGroupSpecification objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

RestrictedInstanceGroupsConfig

The configuration for the restricted instance groups (RIG) in the SageMaker HyperPod cluster.

Type: ClusterRestrictedInstanceGroupsConfig object

Required: No

Tags

Custom tags for managing the SageMaker HyperPod cluster as an AWS resource. You can add tags to your cluster in the same way you add them in other AWS services that support tagging. To learn more about tagging AWS resources in general, see Tagging AWS Resources User Guide.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TieredStorageConfig

The configuration for managed tier checkpointing on the HyperPod cluster. When enabled, this feature uses a multi-tier storage approach for storing model checkpoints, providing faster checkpoint operations and improved fault tolerance across cluster nodes.

Type: ClusterTieredStorageConfig object

Required: No

VpcConfig

Specifies the Amazon Virtual Private Cloud (VPC) that is associated with the Amazon SageMaker HyperPod cluster. You can control access to and from your resources by configuring your VPC. For more information, see Give SageMaker access to resources in your Amazon VPC.

> **Note**
>
> When your Amazon VPC and subnets support IPv6, network communications differ based on the cluster orchestration platform:
>
> - Slurm-orchestrated clusters automatically configure nodes with dual IPv6 and IPv4 addresses, allowing immediate IPv6 network communications.
> - In Amazon EKS-orchestrated clusters, nodes receive dual-stack addressing, but pods can only use IPv6 when the Amazon EKS cluster is explicitly IPv6-enabled. For information about deploying an IPv6 Amazon EKS cluster, see Amazon EKS IPv6 Cluster Deployment.
>
> Additional resources for IPv6 configuration:
>
> - For information about adding IPv6 support to your VPC, see to IPv6 Support for VPC.
> - For information about creating a new IPv6-compatible VPC, see Amazon VPC Creation Guide.
> - To configure SageMaker HyperPod with a custom Amazon VPC, see Custom Amazon VPC Setup for SageMaker HyperPod.

Type: VpcConfig object

Required: No

Response Syntax

```
{
   "ClusterArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ClusterArn

The Amazon Resource Name (ARN) of the cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateClusterSchedulerConfig

Service: Amazon SageMaker Service

Create cluster policy configuration. This policy is used for task prioritization and fair-share allocation of idle compute. This helps prioritize critical workloads and distributes idle compute across entities.

Request Syntax

```
{
   "ClusterArn": "string",
   "Description": "string",
   "Name": "string",
   "SchedulerConfig": {
      "FairShare": "string",
      "IdleResourceSharing": "string",
      "PriorityClasses": [
         {
            "Name": "string",
            "Weight": number
         }
      ]
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterArn

ARN of the cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Required: Yes

Description

Description of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

Name

Name for the cluster policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

SchedulerConfig

Configuration about the monitoring schedule.

Type: SchedulerConfig object

Required: Yes

Tags

Tags of the cluster policy.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "ClusterSchedulerConfigArn": "string",
   "ClusterSchedulerConfigId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ClusterSchedulerConfigArn

ARN of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:cluster-scheduler-config/[a-z0-9]{12}`

ClusterSchedulerConfigId

ID of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateCodeRepository

Service: Amazon SageMaker Service

Creates a Git repository as a resource in your SageMaker AI account. You can associate the repository with notebook instances so that you can use Git source control for the notebooks you create. The Git repository is a resource in your SageMaker AI account, so it can be associated with more than one notebook instance, and it persists independently from the lifecycle of any notebook instances it is associated with.

The repository can be hosted either in AWS CodeCommit or in any other Git repository.

Request Syntax

```
{
   "CodeRepositoryName": "string",
   "GitConfig": {
      "Branch": "string",
      "RepositoryUrl": "string",
      "SecretArn": "string"
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CodeRepositoryName

The name of the Git repository. The name must have 1 to 63 characters. Valid characters are a-z, A-Z, 0-9, and - (hyphen).

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

GitConfig

Specifies details about the repository, including the URL where the repository is located, the default branch, and credentials to use to access the repository.

Type: GitConfig object

Required: Yes

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "CodeRepositoryArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CodeRepositoryArn

The Amazon Resource Name (ARN) of the new repository.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:code-repository/[\S]{1,2048}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateCompilationJob

Service: Amazon SageMaker Service

Starts a model compilation job. After the model has been compiled, Amazon SageMaker AI saves the resulting model artifacts to an Amazon Simple Storage Service (Amazon S3) bucket that you specify.

If you choose to host your model using Amazon SageMaker AI hosting services, you can use the resulting model artifacts as part of the model. You can also use the artifacts with AWS IoT Greengrass. In that case, deploy them as an ML resource.

In the request body, you provide the following:

- A name for the compilation job
- Information about the input model artifacts
- The output location for the compiled model and the device (target) that the model runs on
- The Amazon Resource Name (ARN) of the IAM role that Amazon SageMaker AI assumes to perform the model compilation job.

You can also provide a `Tag` to track the model compilation job's resource use and costs. The response body contains the `CompilationJobArn` for the compiled job.

To stop a model compilation job, use StopCompilationJob. To get information about a particular model compilation job, use DescribeCompilationJob. To get information about multiple model compilation jobs, use ListCompilationJobs.

Request Syntax

```
{
   "CompilationJobName": "string",
   "InputConfig": {
      "DataInputConfig": "string",
      "Framework": "string",
      "FrameworkVersion": "string",
      "S3Uri": "string"
   },
   "ModelPackageVersionArn": "string",
   "OutputConfig": {
      "CompilerOptions": "string",
```

```
      "KmsKeyId": "string",
      "S3OutputLocation": "string",
      "TargetDevice": "string",
      "TargetPlatform": {
         "Accelerator": "string",
         "Arch": "string",
         "Os": "string"
      }
   },
   "RoleArn": "string",
   "StoppingCondition": {
      "MaxPendingTimeInSeconds": number,
      "MaxRuntimeInSeconds": number,
      "MaxWaitTimeInSeconds": number
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "VpcConfig": {
      "SecurityGroupIds": [ "string" ],
      "Subnets": [ "string" ]
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CompilationJobName

A name for the model compilation job. The name must be unique within the AWS Region and within your AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

InputConfig

Provides information about the location of input model artifacts, the name and shape of the expected data inputs, and the framework in which the model was trained.

Type: InputConfig object

Required: No

ModelPackageVersionArn

The Amazon Resource Name (ARN) of a versioned model package. Provide either a `ModelPackageVersionArn` or an `InputConfig` object in the request syntax. The presence of both objects in the `CreateCompilationJob` request will return an exception.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: No

OutputConfig

Provides information about the output location for the compiled model and the target device the model runs on.

Type: OutputConfig object

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of an IAM role that enables Amazon SageMaker AI to perform tasks on your behalf.

During model compilation, Amazon SageMaker AI needs your permission to:

- Read input data from an S3 bucket
- Write model artifacts to an S3 bucket

- Write logs to Amazon CloudWatch Logs
- Publish metrics to Amazon CloudWatch

You grant permissions for all of these tasks to an IAM role. To pass this role to Amazon SageMaker AI, the caller of this API must have the `iam:PassRole` permission. For more information, see Amazon SageMaker AI Roles.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

StoppingCondition

Specifies a limit to how long a model compilation job can run. When the job reaches the time limit, Amazon SageMaker AI ends the compilation job. Use this API to cap model training costs.

Type: StoppingCondition object

Required: Yes

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

VpcConfig

A VpcConfig object that specifies the VPC that you want your compilation job to connect to. Control access to your models by configuring the VPC. For more information, see Protect Compilation Jobs by Using an Amazon Virtual Private Cloud.

Type: NeoVpcConfig object

Required: No

Response Syntax

```
{
   "CompilationJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CompilationJobArn

If the action is successful, the service sends back an HTTP 200 response. Amazon SageMaker AI returns the following data in JSON format:

- `CompilationJobArn`: The Amazon Resource Name (ARN) of the compiled job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:compilation-job/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateComputeQuota

Service: Amazon SageMaker Service

Create compute allocation definition. This defines how compute is allocated, shared, and borrowed for specified entities. Specifically, how to lend and borrow idle compute and assign a fair-share weight to the specified entities.

Request Syntax

```
{
   "ActivationState": "string",
   "ClusterArn": "string",
   "ComputeQuotaConfig": {
      "ComputeQuotaResources": [
         {
            "AcceleratorPartition": {
               "Count": number,
               "Type": "string"
            },
            "Accelerators": number,
            "Count": number,
            "InstanceType": "string",
            "MemoryInGiB": number,
            "VCpu": number
         }
      ],
      "PreemptTeamTasks": "string",
      "ResourceSharingConfig": {
         "AbsoluteBorrowLimits": [
            {
               "AcceleratorPartition": {
                  "Count": number,
                  "Type": "string"
               },
               "Accelerators": number,
               "Count": number,
               "InstanceType": "string",
               "MemoryInGiB": number,
               "VCpu": number
            }
         ],
         "BorrowLimit": number,
         "Strategy": "string"
```

```
      }
   },
   "ComputeQuotaTarget": {
      "FairShareWeight": number,
      "TeamName": "string"
   },
   "Description": "string",
   "Name": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ActivationState

The state of the compute allocation being described. Use to enable or disable compute allocation.

Default is `Enabled`.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

ClusterArn

ARN of the cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Required: Yes

ComputeQuotaConfig

Configuration of the compute allocation definition. This includes the resource sharing option, and the setting to preempt low priority tasks.

Type: ComputeQuotaConfig object

Required: Yes

ComputeQuotaTarget

The target entity to allocate compute resources to.

Type: ComputeQuotaTarget object

Required: Yes

Description

Description of the compute allocation definition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

Name

Name to the compute allocation definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Tags

Tags of the compute allocation definition.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "ComputeQuotaArn": "string",
   "ComputeQuotaId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ComputeQuotaArn

ARN of the compute allocation definition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:compute-quota/[a-z0-9]{12}`

ComputeQuotaId

ID of the compute allocation definition.

Type: String

Pattern: `[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateContext

Service: Amazon SageMaker Service

Creates a *context*. A context is a lineage tracking entity that represents a logical grouping of other tracking or experiment entities. Some examples are an endpoint and a model package. For more information, see Amazon SageMaker ML Lineage Tracking.

Request Syntax

```
{
   "ContextName": "string",
   "ContextType": "string",
   "Description": "string",
   "Properties": {
      "string" : "string"
   },
   "Source": {
      "SourceId": "string",
      "SourceType": "string",
      "SourceUri": "string"
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ContextName

The name of the context. Must be unique to your account in an AWS Region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,119}`

Required: Yes

ContextType

The context type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

Description

The description of the context.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

Properties

A list of properties to add to the context.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 30 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2500.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Required: No

Source

The source type, ID, and URI.

Type: ContextSource object

Required: Yes

Tags

A list of tags to apply to the context.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "ContextArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ContextArn

The Amazon Resource Name (ARN) of the context.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:context/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateDataQualityJobDefinition

Service: Amazon SageMaker Service

Creates a definition for a job that monitors data quality and drift. For information about model monitor, see Amazon SageMaker AI Model Monitor.

Request Syntax

```
{
   "DataQualityAppSpecification": {
      "ContainerArguments": [ "string" ],
      "ContainerEntrypoint": [ "string" ],
      "Environment": {
         "string" : "string"
      },
      "ImageUri": "string",
      "PostAnalyticsProcessorSourceUri": "string",
      "RecordPreprocessorSourceUri": "string"
   },
   "DataQualityBaselineConfig": {
      "BaseliningJobName": "string",
      "ConstraintsResource": {
         "S3Uri": "string"
      },
      "StatisticsResource": {
         "S3Uri": "string"
      }
   },
   "DataQualityJobInput": {
      "BatchTransformInput": {
         "DataCapturedDestinationS3Uri": "string",
         "DatasetFormat": {
            "Csv": {
               "Header": boolean
            },
            "Json": {
               "Line": boolean
            },
            "Parquet": {
            }
         },
         "EndTimeOffset": "string",
         "ExcludeFeaturesAttribute": "string",
```

```
            "FeaturesAttribute": "string",
            "InferenceAttribute": "string",
            "LocalPath": "string",
            "ProbabilityAttribute": "string",
            "ProbabilityThresholdAttribute": number,
            "S3DataDistributionType": "string",
            "S3InputMode": "string",
            "StartTimeOffset": "string"
         },
         "EndpointInput": {
            "EndpointName": "string",
            "EndTimeOffset": "string",
            "ExcludeFeaturesAttribute": "string",
            "FeaturesAttribute": "string",
            "InferenceAttribute": "string",
            "LocalPath": "string",
            "ProbabilityAttribute": "string",
            "ProbabilityThresholdAttribute": number,
            "S3DataDistributionType": "string",
            "S3InputMode": "string",
            "StartTimeOffset": "string"
         }
      },
      "DataQualityJobOutputConfig": {
         "KmsKeyId": "string",
         "MonitoringOutputs": [
            {
               "S3Output": {
                  "LocalPath": "string",
                  "S3UploadMode": "string",
                  "S3Uri": "string"
               }
            }
         ]
      },
      "JobDefinitionName": "string",
      "JobResources": {
         "ClusterConfig": {
            "InstanceCount": number,
            "InstanceType": "string",
            "VolumeKmsKeyId": "string",
            "VolumeSizeInGB": number
         }
      },
```

```
   "NetworkConfig": {
      "EnableInterContainerTrafficEncryption": boolean,
      "EnableNetworkIsolation": boolean,
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
   },
   "RoleArn": "string",
   "StoppingCondition": {
      "MaxRuntimeInSeconds": number
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DataQualityAppSpecification

Specifies the container that runs the monitoring job.

Type: DataQualityAppSpecification object

Required: Yes

DataQualityBaselineConfig

Configures the constraints and baselines for the monitoring job.

Type: DataQualityBaselineConfig object

Required: No

DataQualityJobInput

A list of inputs for the monitoring job. Currently endpoints are supported as monitoring inputs.

Type: DataQualityJobInput object

Required: Yes

DataQualityJobOutputConfig

The output configuration for monitoring jobs.

Type: MonitoringOutputConfig object

Required: Yes

JobDefinitionName

The name for the monitoring job definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

JobResources

Identifies the resources to deploy for a monitoring job.

Type: MonitoringResources object

Required: Yes

NetworkConfig

Specifies networking configuration for the monitoring job.

Type: MonitoringNetworkConfig object

Required: No

RoleArn

The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker AI can assume to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

StoppingCondition

A time limit for how long the monitoring job is allowed to run before stopping.

Type: MonitoringStoppingCondition object

Required: No

Tags

(Optional) An array of key-value pairs. For more information, see Using Cost Allocation Tags in the *AWS Billing and Cost Management User Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "JobDefinitionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobDefinitionArn

The Amazon Resource Name (ARN) of the job definition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateDeviceFleet

Service: Amazon SageMaker Service

Creates a device fleet.

Request Syntax

```
{
   "Description": "string",
   "DeviceFleetName": "string",
   "EnableIotRoleAlias": boolean,
   "OutputConfig": {
      "KmsKeyId": "string",
      "PresetDeploymentConfig": "string",
      "PresetDeploymentType": "string",
      "S3OutputLocation": "string"
   },
   "RoleArn": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Description

A description of the fleet.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 800.

Pattern: `[-a-zA-Z0-9_.,;:! ]*`

Required: No

DeviceFleetName

The name of the fleet that the device belongs to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EnableIotRoleAlias

Whether to create an AWS IoT Role Alias during device fleet creation. The name of the role alias generated will match this pattern: "SageMakerEdge-{DeviceFleetName}".

For example, if your device fleet is called "demo-fleet", the name of the role alias will be "SageMakerEdge-demo-fleet".

Type: Boolean

Required: No

OutputConfig

The output configuration for storing sample data collected by the fleet.

Type: EdgeOutputConfig object

Required: Yes

RoleArn

The Amazon Resource Name (ARN) that has access to AWS Internet of Things (IoT).

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

Tags

Creates tags for the specified fleet.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateDomain

Service: Amazon SageMaker Service

Creates a `Domain`. A domain consists of an associated Amazon Elastic File System volume, a list of authorized users, and a variety of security, application, policy, and Amazon Virtual Private Cloud (VPC) configurations. Users within a domain can share notebook files and other artifacts with each other.

EFS storage

When a domain is created, an EFS volume is created for use by all of the users within the domain. Each user receives a private home directory within the EFS volume for notebooks, Git repositories, and data files.

SageMaker AI uses the AWS Key Management Service (AWS KMS) to encrypt the EFS volume attached to the domain with an AWS managed key by default. For more control, you can specify a customer managed key. For more information, see Protect Data at Rest Using Encryption.

VPC configuration

All traffic between the domain and the Amazon EFS volume is through the specified VPC and subnets. For other traffic, you can specify the `AppNetworkAccessType` parameter. `AppNetworkAccessType` corresponds to the network access type that you choose when you onboard to the domain. The following options are available:

- `PublicInternetOnly` - Non-EFS traffic goes through a VPC managed by Amazon SageMaker AI, which allows internet access. This is the default value.
- `VpcOnly` - All traffic is through the specified VPC and subnets. Internet access is disabled by default. To allow internet access, you must specify a NAT gateway.

  When internet access is disabled, you won't be able to run a Amazon SageMaker AI Studio notebook or to train or host models unless your VPC has an interface endpoint to the SageMaker AI API and runtime or a NAT gateway and your security groups allow outbound connections.

> **Important**
>
> NFS traffic over TCP on port 2049 needs to be allowed in both inbound and outbound rules in order to launch a Amazon SageMaker AI Studio app successfully.

For more information, see Connect Amazon SageMaker AI Studio Notebooks to Resources in a VPC.

Request Syntax

```
{
   "AppNetworkAccessType": "string",
   "AppSecurityGroupManagement": "string",
   "AuthMode": "string",
   "DefaultSpaceSettings": {
      "CustomFileSystemConfigs": [
         { ... }
      ],
      "CustomPosixUserConfig": {
         "Gid": number,
         "Uid": number
      },
      "ExecutionRole": "string",
      "JupyterLabAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number,
               "LifecycleManagement": "string",
               "MaxIdleTimeoutInMinutes": number,
               "MinIdleTimeoutInMinutes": number
            }
         },
         "BuiltInLifecycleConfigArn": "string",
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
```

```
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "EmrSettings": {
            "AssumableRoleArns": [ "string" ],
            "ExecutionRoleArns": [ "string" ]
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "JupyterServerAppSettings": {
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "KernelGatewayAppSettings": {
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "SecurityGroups": [ "string" ],
```

```
      "SpaceStorageSettings": {
         "DefaultEbsStorageSettings": {
            "DefaultEbsVolumeSizeInGb": number,
            "MaximumEbsVolumeSizeInGb": number
         }
      }
   },
   "DefaultUserSettings": {
      "AutoMountHomeEFS": "string",
      "CanvasAppSettings": {
         "DirectDeploySettings": {
            "Status": "string"
         },
         "EmrServerlessSettings": {
            "ExecutionRoleArn": "string",
            "Status": "string"
         },
         "GenerativeAiSettings": {
            "AmazonBedrockRoleArn": "string"
         },
         "IdentityProviderOAuthSettings": [
            {
               "DataSourceName": "string",
               "SecretArn": "string",
               "Status": "string"
            }
         ],
         "KendraSettings": {
            "Status": "string"
         },
         "ModelRegisterSettings": {
            "CrossAccountModelRegisterRoleArn": "string",
            "Status": "string"
         },
         "TimeSeriesForecastingSettings": {
            "AmazonForecastRoleArn": "string",
            "Status": "string"
         },
         "WorkspaceSettings": {
            "S3ArtifactPath": "string",
            "S3KmsKeyId": "string"
         }
      },
      "CodeEditorAppSettings": {
```

```
            "AppLifecycleManagement": {
               "IdleSettings": {
                  "IdleTimeoutInMinutes": number,
                  "LifecycleManagement": "string",
                  "MaxIdleTimeoutInMinutes": number,
                  "MinIdleTimeoutInMinutes": number
               }
            },
            "BuiltInLifecycleConfigArn": "string",
            "CustomImages": [
               {
                  "AppImageConfigName": "string",
                  "ImageName": "string",
                  "ImageVersionNumber": number
               }
            ],
            "DefaultResourceSpec": {
               "InstanceType": "string",
               "LifecycleConfigArn": "string",
               "SageMakerImageArn": "string",
               "SageMakerImageVersionAlias": "string",
               "SageMakerImageVersionArn": "string",
               "TrainingPlanArn": "string"
            },
            "LifecycleConfigArns": [ "string" ]
         },
         "CustomFileSystemConfigs": [
            { ... }
         ],
         "CustomPosixUserConfig": {
            "Gid": number,
            "Uid": number
         },
         "DefaultLandingUri": "string",
         "ExecutionRole": "string",
         "JupyterLabAppSettings": {
            "AppLifecycleManagement": {
               "IdleSettings": {
                  "IdleTimeoutInMinutes": number,
                  "LifecycleManagement": "string",
                  "MaxIdleTimeoutInMinutes": number,
                  "MinIdleTimeoutInMinutes": number
               }
            },
```

```
            "BuiltInLifecycleConfigArn": "string",
            "CodeRepositories": [
               {
                  "RepositoryUrl": "string"
               }
            ],
            "CustomImages": [
               {
                  "AppImageConfigName": "string",
                  "ImageName": "string",
                  "ImageVersionNumber": number
               }
            ],
            "DefaultResourceSpec": {
               "InstanceType": "string",
               "LifecycleConfigArn": "string",
               "SageMakerImageArn": "string",
               "SageMakerImageVersionAlias": "string",
               "SageMakerImageVersionArn": "string",
               "TrainingPlanArn": "string"
            },
            "EmrSettings": {
               "AssumableRoleArns": [ "string" ],
               "ExecutionRoleArns": [ "string" ]
            },
            "LifecycleConfigArns": [ "string" ]
         },
         "JupyterServerAppSettings": {
            "CodeRepositories": [
               {
                  "RepositoryUrl": "string"
               }
            ],
            "DefaultResourceSpec": {
               "InstanceType": "string",
               "LifecycleConfigArn": "string",
               "SageMakerImageArn": "string",
               "SageMakerImageVersionAlias": "string",
               "SageMakerImageVersionArn": "string",
               "TrainingPlanArn": "string"
            },
            "LifecycleConfigArns": [ "string" ]
         },
         "KernelGatewayAppSettings": {
```

```
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "RSessionAppSettings": {
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         }
      },
      "RStudioServerProAppSettings": {
         "AccessStatus": "string",
         "UserGroup": "string"
      },
      "SecurityGroups": [ "string" ],
      "SharingSettings": {
         "NotebookOutputOption": "string",
         "S3KmsKeyId": "string",
         "S3OutputPath": "string"
      },
```

```
      "SpaceStorageSettings": {
         "DefaultEbsStorageSettings": {
            "DefaultEbsVolumeSizeInGb": number,
            "MaximumEbsVolumeSizeInGb": number
         }
      },
      "StudioWebPortal": "string",
      "StudioWebPortalSettings": {
         "ExecutionRoleSessionNameMode": "string",
         "HiddenAppTypes": [ "string" ],
         "HiddenInstanceTypes": [ "string" ],
         "HiddenMlTools": [ "string" ],
         "HiddenSageMakerImageVersionAliases": [
            {
               "SageMakerImageName": "string",
               "VersionAliases": [ "string" ]
            }
         ]
      },
      "TensorBoardAppSettings": {
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         }
      }
   },
   "DomainName": "string",
   "DomainSettings": {
      "AmazonQSettings": {
         "QProfileArn": "string",
         "Status": "string"
      },
      "DockerSettings": {
         "EnableDockerAccess": "string",
         "RootlessDocker": "string",
         "VpcOnlyTrustedAccounts": [ "string" ]
      },
      "ExecutionRoleIdentityConfig": "string",
      "IpAddressType": "string",
      "RStudioServerProDomainSettings": {
```

```
            "DefaultResourceSpec": {
               "InstanceType": "string",
               "LifecycleConfigArn": "string",
               "SageMakerImageArn": "string",
               "SageMakerImageVersionAlias": "string",
               "SageMakerImageVersionArn": "string",
               "TrainingPlanArn": "string"
            },
            "DomainExecutionRoleArn": "string",
            "RStudioConnectUrl": "string",
            "RStudioPackageManagerUrl": "string"
         },
         "SecurityGroupIds": [ "string" ],
         "TrustedIdentityPropagationSettings": {
            "Status": "string"
         },
         "UnifiedStudioSettings": {
            "DomainAccountId": "string",
            "DomainId": "string",
            "DomainRegion": "string",
            "EnvironmentId": "string",
            "ProjectId": "string",
            "ProjectS3Path": "string",
            "SingleSignOnApplicationArn": "string",
            "StudioWebPortalAccess": "string"
         }
      },
      "HomeEfsFileSystemCreation": "string",
      "HomeEfsFileSystemKmsKeyId": "string",
      "KmsKeyId": "string",
      "SubnetIds": [ "string" ],
      "TagPropagation": "string",
      "Tags": [
         {
            "Key": "string",
            "Value": "string"
         }
      ],
      "VpcId": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AppNetworkAccessType

Specifies the VPC used for non-EFS traffic. The default value is `PublicInternetOnly`.

- `PublicInternetOnly` - Non-EFS traffic is through a VPC managed by Amazon SageMaker AI, which allows direct internet access
- `VpcOnly` - All traffic is through the specified VPC and subnets

Type: String

Valid Values: `PublicInternetOnly | VpcOnly`

Required: No

AppSecurityGroupManagement

The entity that creates and manages the required security groups for inter-app communication in `VPCOnly` mode. Required when `CreateDomain.AppNetworkAccessType` is `VPCOnly` and `DomainSettings.RStudioServerProDomainSettings.DomainExecutionRoleArn` is provided. If setting up the domain for use with RStudio, this value must be set to `Service`.

Type: String

Valid Values: `Service | Customer`

Required: No

AuthMode

The mode of authentication that members use to access the domain.

Type: String

Valid Values: `SSO | IAM`

Required: Yes

DefaultSpaceSettings

The default settings for shared spaces that users create in the domain.

Type: DefaultSpaceSettings object

Required: No

DefaultUserSettings

The default settings to use to create a user profile when `UserSettings` isn't specified in the call to the `CreateUserProfile` API.

`SecurityGroups` is aggregated when specified in both calls. For all other settings in `UserSettings`, the values specified in `CreateUserProfile` take precedence over those specified in `CreateDomain`.

Type: UserSettings object

Required: Yes

DomainName

A name for the domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

DomainSettings

A collection of `Domain` settings.

Type: DomainSettings object

Required: No

HomeEfsFileSystemCreation

Indicates whether to create a home EFS file system for the domain. Defaults to `Enabled`. Set to `Disabled` to skip EFS creation and reduce domain creation time. You can enable EFS later by calling `UpdateDomain`.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

HomeEfsFileSystemKmsKeyId

This parameter has been deprecated.

Use `KmsKeyId`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

KmsKeyId

SageMaker AI uses AWS KMS to encrypt EFS and EBS volumes attached to the domain with an AWS managed key by default. For more control, specify a customer managed key.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

SubnetIds

The VPC subnets that the domain uses for communication.

The field is optional when the `AppNetworkAccessType` parameter is set to `PublicInternetOnly` for domains created from Amazon SageMaker Unified Studio.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: No

TagPropagation

Indicates whether custom tag propagation is supported for the domain. Defaults to DISABLED.

Type: String

Valid Values: ENABLED | DISABLED

Required: No

Tags

Tags to associated with the Domain. Each tag consists of a key and an optional value. Tag keys must be unique per resource. Tags are searchable using the Search API.

Tags that you specify for the Domain are also added to all Apps that the Domain launches.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

VpcId

The ID of the Amazon Virtual Private Cloud (VPC) that the domain uses for communication.

The field is optional when the AppNetworkAccessType parameter is set to PublicInternetOnly for domains created from Amazon SageMaker Unified Studio.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: [-0-9a-zA-Z]+

Required: No

Response Syntax

```
{
   "DomainArn": "string",
   "DomainId": "string",
```

```
   "Url": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

DomainArn

The Amazon Resource Name (ARN) of the created domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:domain/.*`

DomainId

The ID of the created domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Url

The URL to the created domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateEdgeDeploymentPlan

Service: Amazon SageMaker Service

Creates an edge deployment plan, consisting of multiple stages. Each stage may have a different deployment configuration and devices.

Request Syntax

```
{
   "DeviceFleetName": "string",
   "EdgeDeploymentPlanName": "string",
   "ModelConfigs": [
      {
         "EdgePackagingJobName": "string",
         "ModelHandle": "string"
      }
   ],
   "Stages": [
      {
         "DeploymentConfig": {
            "FailureHandlingPolicy": "string"
         },
         "DeviceSelectionConfig": {
            "DeviceNameContains": "string",
            "DeviceNames": [ "string" ],
            "DeviceSubsetType": "string",
            "Percentage": number
         },
         "StageName": "string"
      }
   ],
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeviceFleetName

The device fleet used for this edge deployment plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EdgeDeploymentPlanName

The name of the edge deployment plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelConfigs

List of models associated with the edge deployment plan.

Type: Array of EdgeDeploymentModelConfig objects

Required: Yes

Stages

List of stages of the edge deployment plan. The number of stages is limited to 10 per deployment.

Type: Array of DeploymentStage objects

Required: No

Tags

List of tags with which to tag the edge deployment plan.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "EdgeDeploymentPlanArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

EdgeDeploymentPlanArn

The ARN of the edge deployment plan.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z\-]*:\d{12}:edge-deployment/?[a-zA-Z_0-9+=,.@\-_/]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateEdgeDeploymentStage

Service: Amazon SageMaker Service

Creates a new stage in an existing edge deployment plan.

Request Syntax

```
{
   "EdgeDeploymentPlanName": "string",
   "Stages": [
      {
         "DeploymentConfig": {
            "FailureHandlingPolicy": "string"
         },
         "DeviceSelectionConfig": {
            "DeviceNameContains": "string",
            "DeviceNames": [ "string" ],
            "DeviceSubsetType": "string",
            "Percentage": number
         },
         "StageName": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EdgeDeploymentPlanName

The name of the edge deployment plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Stages

List of stages to be added to the edge deployment plan.

Type: Array of DeploymentStage objects

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateEdgePackagingJob

Service: Amazon SageMaker Service

Starts a SageMaker Edge Manager model packaging job. Edge Manager will use the model artifacts from the Amazon Simple Storage Service bucket that you specify. After the model has been packaged, Amazon SageMaker saves the resulting artifacts to an S3 bucket that you specify.

Request Syntax

```
{
   "CompilationJobName": "string",
   "EdgePackagingJobName": "string",
   "ModelName": "string",
   "ModelVersion": "string",
   "OutputConfig": {
      "KmsKeyId": "string",
      "PresetDeploymentConfig": "string",
      "PresetDeploymentType": "string",
      "S3OutputLocation": "string"
   },
   "ResourceKey": "string",
   "RoleArn": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CompilationJobName

The name of the SageMaker Neo compilation job that will be used to locate model artifacts for packaging.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EdgePackagingJobName

The name of the edge packaging job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelName

The name of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelVersion

The version of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 30.

Pattern: `[a-zA-Z0-9\ \_\.]+`

Required: Yes

OutputConfig

Provides information about the output location for the packaged model.

Type: EdgeOutputConfig object

Required: Yes

ResourceKey

The AWS KMS key to use when encrypting the EBS volume the edge packaging job runs on.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

RoleArn

The Amazon Resource Name (ARN) of an IAM role that enables Amazon SageMaker to download and upload the model, and to contact SageMaker Neo.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Tags

Creates tags for the packaging job.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateEndpoint

Service: Amazon SageMaker Service

Creates an endpoint using the endpoint configuration specified in the request. SageMaker uses the endpoint to provision resources and deploy models. You create the endpoint configuration with the CreateEndpointConfig API.

Use this API to deploy models using SageMaker hosting services.

> **Note**
>
> You must not delete an `EndpointConfig` that is in use by an endpoint that is live or while the `UpdateEndpoint` or `CreateEndpoint` operations are being performed on the endpoint. To update an endpoint, you must create a new `EndpointConfig`.

The endpoint name must be unique within an AWS Region in your AWS account.

When it receives the request, SageMaker creates the endpoint, launches the resources (ML compute instances), and deploys the model(s) on them.

> **Note**
>
> When you call CreateEndpoint, a load call is made to DynamoDB to verify that your endpoint configuration exists. When you read data from a DynamoDB table supporting `Eventually Consistent Reads`, the response might not reflect the results of a recently completed write operation. The response might include some stale data. If the dependent entities are not yet in DynamoDB, this causes a validation error. If you repeat your read request after a short time, the response should return the latest data. So retry logic is recommended to handle these possible issues. We also recommend that customers call DescribeEndpointConfig before calling CreateEndpoint to minimize the potential impact of a DynamoDB eventually consistent read.

When SageMaker receives the request, it sets the endpoint status to `Creating`. After it creates the endpoint, it sets the status to `InService`. SageMaker can then process incoming requests for inferences. To check the status of an endpoint, use the DescribeEndpoint API.

If any of the models hosted at this endpoint get model data from an Amazon S3 location, SageMaker uses AWS Security Token Service to download model artifacts from the S3 path you provided. AWS STS is activated in your AWS account by default. If you previously deactivated AWS STS for a region, you need to reactivate AWS STS for that region. For more information, see Activating and Deactivating AWS STS in an AWS Region in the *AWS Identity and Access Management User Guide*.

Note

To add the IAM role policies for using this API operation, go to the IAM console, and choose Roles in the left navigation pane. Search the IAM role that you want to grant access to use the CreateEndpoint and CreateEndpointConfig API operations, add the following policies to the role.

- Option 1: For a full SageMaker access, search and attach the `AmazonSageMakerFullAccess` policy.
- Option 2: For granting a limited access to an IAM role, paste the following Action elements manually into the JSON file of the IAM role:

```
"Action": ["sagemaker:CreateEndpoint",
"sagemaker:CreateEndpointConfig"]

"Resource": [

"arn:aws:sagemaker:region:account-id:endpoint/endpointName"

"arn:aws:sagemaker:region:account-id:endpoint-config/
endpointConfigName"

]
```

For more information, see SageMaker API Permissions: Actions, Permissions, and Resources Reference.

Request Syntax

```
{
   "DeploymentConfig": {
      "AutoRollbackConfiguration": {
```

```
         "Alarms": [
            {
               "AlarmName": "string"
            }
         ]
      },
      "BlueGreenUpdatePolicy": {
         "MaximumExecutionTimeoutInSeconds": number,
         "TerminationWaitInSeconds": number,
         "TrafficRoutingConfiguration": {
            "CanarySize": {
               "Type": "string",
               "Value": number
            },
            "LinearStepSize": {
               "Type": "string",
               "Value": number
            },
            "Type": "string",
            "WaitIntervalInSeconds": number
         }
      },
      "RollingUpdatePolicy": {
         "MaximumBatchSize": {
            "Type": "string",
            "Value": number
         },
         "MaximumExecutionTimeoutInSeconds": number,
         "RollbackMaximumBatchSize": {
            "Type": "string",
            "Value": number
         },
         "WaitIntervalInSeconds": number
      }
   },
   "EndpointConfigName": "string",
   "EndpointName": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
```

```
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeploymentConfig

The deployment configuration for an endpoint, which contains the desired deployment strategy and rollback configurations.

Type: DeploymentConfig object

Required: No

EndpointConfigName

The name of an endpoint configuration. For more information, see CreateEndpointConfig.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EndpointName

The name of the endpoint.The name must be unique within an AWS Region in your AWS account. The name is case-insensitive in `CreateEndpoint`, but the case is preserved and must be matched in InvokeEndpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "EndpointArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

EndpointArn

The Amazon Resource Name (ARN) of the endpoint.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateEndpointConfig

Service: Amazon SageMaker Service

Creates an endpoint configuration that SageMaker hosting services uses to deploy models. In the configuration, you identify one or more models, created using the `CreateModel` API, to deploy and the resources that you want SageMaker to provision. Then you call the CreateEndpoint API.

> **Note**
>
> Use this API if you want to use SageMaker hosting services to deploy models into production.

In the request, you define a `ProductionVariant`, for each model that you want to deploy. Each `ProductionVariant` parameter also describes the resources that you want SageMaker to provision. This includes the number and type of ML compute instances to deploy.

If you are hosting multiple models, you also assign a `VariantWeight` to specify how much traffic you want to allocate to each model. For example, suppose that you want to host two models, A and B, and you assign traffic weight 2 for model A and 1 for model B. SageMaker distributes two-thirds of the traffic to Model A, and one-third to model B.

> **Note**
>
> When you call CreateEndpoint, a load call is made to DynamoDB to verify that your endpoint configuration exists. When you read data from a DynamoDB table supporting `Eventually Consistent Reads`, the response might not reflect the results of a recently completed write operation. The response might include some stale data. If the dependent entities are not yet in DynamoDB, this causes a validation error. If you repeat your read request after a short time, the response should return the latest data. So retry logic is recommended to handle these possible issues. We also recommend that customers call DescribeEndpointConfig before calling CreateEndpoint to minimize the potential impact of a DynamoDB eventually consistent read.

Request Syntax

```
{
```

```
   "AsyncInferenceConfig": {
      "ClientConfig": {
         "MaxConcurrentInvocationsPerInstance": number
      },
      "OutputConfig": {
         "KmsKeyId": "string",
         "NotificationConfig": {
            "ErrorTopic": "string",
            "IncludeInferenceResponseIn": [ "string" ],
            "SuccessTopic": "string"
         },
         "S3FailurePath": "string",
         "S3OutputPath": "string"
      }
   },
   "DataCaptureConfig": {
      "CaptureContentTypeHeader": {
         "CsvContentTypes": [ "string" ],
         "JsonContentTypes": [ "string" ]
      },
      "CaptureOptions": [
         {
            "CaptureMode": "string"
         }
      ],
      "DestinationS3Uri": "string",
      "EnableCapture": boolean,
      "InitialSamplingPercentage": number,
      "KmsKeyId": "string"
   },
   "EnableNetworkIsolation": boolean,
   "EndpointConfigName": "string",
   "ExecutionRoleArn": "string",
   "ExplainerConfig": {
      "ClarifyExplainerConfig": {
         "EnableExplanations": "string",
         "InferenceConfig": {
            "ContentTemplate": "string",
            "FeatureHeaders": [ "string" ],
            "FeaturesAttribute": "string",
            "FeatureTypes": [ "string" ],
            "LabelAttribute": "string",
            "LabelHeaders": [ "string" ],
            "LabelIndex": number,
```

```
            "MaxPayloadInMB": number,
            "MaxRecordCount": number,
            "ProbabilityAttribute": "string",
            "ProbabilityIndex": number
         },
         "ShapConfig": {
            "NumberOfSamples": number,
            "Seed": number,
            "ShapBaselineConfig": {
               "MimeType": "string",
               "ShapBaseline": "string",
               "ShapBaselineUri": "string"
            },
            "TextConfig": {
               "Granularity": "string",
               "Language": "string"
            },
            "UseLogit": boolean
         }
      }
   },
   "KmsKeyId": "string",
   "MetricsConfig": {
      "EnableDetailedObservability": boolean,
      "EnableEnhancedMetrics": boolean,
      "MetricPublishFrequencyInSeconds": number
   },
   "ProductionVariants": [
      {
         "AcceleratorType": "string",
         "CapacityReservationConfig": {
            "CapacityReservationPreference": "string",
            "MlReservationArn": "string"
         },
         "ContainerStartupHealthCheckTimeoutInSeconds": number,
         "CoreDumpConfig": {
            "DestinationS3Uri": "string",
            "KmsKeyId": "string"
         },
         "EnableSSMAccess": boolean,
         "InferenceAmiVersion": "string",
         "InitialInstanceCount": number,
         "InitialVariantWeight": number,
         "InstancePools": [
```

```
            {
               "InstanceType": "string",
               "ModelNameOverride": "string",
               "Priority": number
            }
         ],
         "InstanceType": "string",
         "ManagedInstanceScaling": {
            "MaxInstanceCount": number,
            "MinInstanceCount": number,
            "ScaleInPolicy": {
               "CooldownInMinutes": number,
               "MaximumStepSize": number,
               "Strategy": "string"
            },
            "Status": "string"
         },
         "ModelDataDownloadTimeoutInSeconds": number,
         "ModelName": "string",
         "RoutingConfig": {
            "RoutingStrategy": "string"
         },
         "ServerlessConfig": {
            "MaxConcurrency": number,
            "MemorySizeInMB": number,
            "ProvisionedConcurrency": number
         },
         "VariantInstanceProvisionTimeoutInSeconds": number,
         "VariantName": "string",
         "VolumeSizeInGB": number
      }
   ],
   "ShadowProductionVariants": [
      {
         "AcceleratorType": "string",
         "CapacityReservationConfig": {
            "CapacityReservationPreference": "string",
            "MlReservationArn": "string"
         },
         "ContainerStartupHealthCheckTimeoutInSeconds": number,
         "CoreDumpConfig": {
            "DestinationS3Uri": "string",
            "KmsKeyId": "string"
         },
```

```
         "EnableSSMAccess": boolean,
         "InferenceAmiVersion": "string",
         "InitialInstanceCount": number,
         "InitialVariantWeight": number,
         "InstancePools": [
            {
               "InstanceType": "string",
               "ModelNameOverride": "string",
               "Priority": number
            }
         ],
         "InstanceType": "string",
         "ManagedInstanceScaling": {
            "MaxInstanceCount": number,
            "MinInstanceCount": number,
            "ScaleInPolicy": {
               "CooldownInMinutes": number,
               "MaximumStepSize": number,
               "Strategy": "string"
            },
            "Status": "string"
         },
         "ModelDataDownloadTimeoutInSeconds": number,
         "ModelName": "string",
         "RoutingConfig": {
            "RoutingStrategy": "string"
         },
         "ServerlessConfig": {
            "MaxConcurrency": number,
            "MemorySizeInMB": number,
            "ProvisionedConcurrency": number
         },
         "VariantInstanceProvisionTimeoutInSeconds": number,
         "VariantName": "string",
         "VolumeSizeInGB": number
      }
   ],
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "VpcConfig": {
```

```
        "SecurityGroupIds": [ "string" ],
        "Subnets": [ "string" ]
    }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AsyncInferenceConfig

Specifies configuration for how an endpoint performs asynchronous inference. This is a required field in order for your Endpoint to be invoked using InvokeEndpointAsync.

Type: AsyncInferenceConfig object

Required: No

DataCaptureConfig

Configuration to control how SageMaker AI captures inference data.

Type: DataCaptureConfig object

Required: No

EnableNetworkIsolation

Sets whether all model containers deployed to the endpoint are isolated. If they are, no inbound or outbound network calls can be made to or from the model containers.

Type: Boolean

Required: No

EndpointConfigName

The name of the endpoint configuration. You specify this name in a CreateEndpoint request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ExecutionRoleArn

The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker AI can assume to perform actions on your behalf. For more information, see SageMaker AI Roles.

> **Note**
>
> To be able to pass this role to Amazon SageMaker AI, the caller of this action must have the `iam:PassRole` permission.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

ExplainerConfig

A member of `CreateEndpointConfig` that enables explainers.

Type: ExplainerConfig object

Required: No

KmsKeyId

The Amazon Resource Name (ARN) of a AWS Key Management Service key that SageMaker uses to encrypt data on the storage volume attached to the ML compute instance that hosts the endpoint.

The KmsKeyId can be any of the following formats:

- Key ID: `1234abcd-12ab-34cd-56ef-1234567890ab`
- Key ARN: `arn:aws:kms:us-west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab`
- Alias name: `alias/ExampleAlias`

- Alias name ARN: `arn:aws:kms:us-west-2:111122223333:alias/ExampleAlias`

The KMS key policy must grant permission to the IAM role that you specify in your `CreateEndpoint`, `UpdateEndpoint` requests. For more information, refer to the AWS Key Management Service section Using Key Policies in AWS KMS

> **Note**
>
> Certain Nitro-based instances include local storage, dependent on the instance type. Local storage volumes are encrypted using a hardware module on the instance. If any of the models that you specify in the `ProductionVariants` parameter use nitro-based instances with local storage, the `KmsKeyId` parameter does not encrypt instance local storage.
> For a list of instance types that support local instance storage, see Instance Store Volumes.
> For more information about local instance storage encryption, see SSD Instance Store Volumes.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

MetricsConfig

The configuration parameters for utilization metrics.

Type: MetricsConfig object

Required: No

ProductionVariants

An array of `ProductionVariant` objects, one for each model that you want to host at this endpoint.

Type: Array of ProductionVariant objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: Yes

ShadowProductionVariants

An array of `ProductionVariant` objects, one for each model that you want to host at this endpoint in shadow mode with production traffic replicated from the model specified on `ProductionVariants`. If you use this field, you can only specify one variant for `ProductionVariants` and one variant for `ShadowProductionVariants`.

Type: Array of ProductionVariant objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

VpcConfig

Specifies an Amazon Virtual Private Cloud (VPC) that your SageMaker jobs, hosted models, and compute resources have access to. You can control access to and from your resources by configuring a VPC. For more information, see Give SageMaker Access to Resources in your Amazon VPC.

Type: VpcConfig object

Required: No

Response Syntax

```
{
```

```
   "EndpointConfigArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

EndpointConfigArn

The Amazon Resource Name (ARN) of the endpoint configuration.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint-config/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateExperiment

Service: Amazon SageMaker Service

Creates a SageMaker *experiment*. An experiment is a collection of *trials* that are observed, compared and evaluated as a group. A trial is a set of steps, called *trial components*, that produce a machine learning model.

> **Note**
>
> In the Studio UI, trials are referred to as *run groups* and trial components are referred to as *runs*.

The goal of an experiment is to determine the components that produce the best model. Multiple trials are performed, each one isolating and measuring the impact of a change to one or more inputs, while keeping the remaining inputs constant.

When you use SageMaker Studio or the SageMaker Python SDK, all experiments, trials, and trial components are automatically tracked, logged, and indexed. When you use the AWS SDK for Python (Boto), you must use the logging APIs provided by the SDK.

You can add tags to experiments, trials, trial components and then use the Search API to search for the tags.

To add a description to an experiment, specify the optional `Description` parameter. To add a description later, or to change the description, call the UpdateExperiment API.

To get a list of all your experiments, call the ListExperiments API. To view an experiment's properties, call the DescribeExperiment API. To get a list of all the trials associated with an experiment, call the ListTrials API. To create a trial call the CreateTrial API.

Request Syntax

```
{
   "Description": "string",
   "DisplayName": "string",
   "ExperimentName": "string",
   "Tags": [
      {
         "Key": "string",
```

```
            "Value": "string"
        }
    ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Description

The description of the experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

DisplayName

The name of the experiment as displayed. The name doesn't need to be unique. If you don't specify `DisplayName`, the value in `ExperimentName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

ExperimentName

The name of the experiment. The name must be unique in your AWS account and is not case-sensitive.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Tags

A list of tags to associate with the experiment. You can use Search API to search on the tags.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "ExperimentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ExperimentArn

The Amazon Resource Name (ARN) of the experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateFeatureGroup

Service: Amazon SageMaker Service

Create a new `FeatureGroup`. A `FeatureGroup` is a group of `Features` defined in the `FeatureStore` to describe a `Record`.

The `FeatureGroup` defines the schema and features contained in the `FeatureGroup`. A `FeatureGroup` definition is composed of a list of `Features`, a `RecordIdentifierFeatureName`, an `EventTimeFeatureName` and configurations for its `OnlineStore` and `OfflineStore`. Check AWS service quotas to see the `FeatureGroups` quota for your AWS account.

Note that it can take approximately 10-15 minutes to provision an `OnlineStore` `FeatureGroup` with the `InMemory` `StorageType`.

> **Important**
>
> You must include at least one of `OnlineStoreConfig` and `OfflineStoreConfig` to create a `FeatureGroup`.

Request Syntax

```
{
   "Description": "string",
   "EventTimeFeatureName": "string",
   "FeatureDefinitions": [
      {
         "CollectionConfig": { ... },
         "CollectionType": "string",
         "FeatureName": "string",
         "FeatureType": "string"
      }
   ],
   "FeatureGroupName": "string",
   "OfflineStoreConfig": {
      "DataCatalogConfig": {
         "Catalog": "string",
         "Database": "string",
         "TableName": "string"
```

```
      },
      "DisableGlueTableCreation": boolean,
      "S3StorageConfig": {
         "KmsKeyId": "string",
         "ResolvedOutputS3Uri": "string",
         "S3Uri": "string"
      },
      "TableFormat": "string"
   },
   "OnlineStoreConfig": {
      "EnableOnlineStore": boolean,
      "SecurityConfig": {
         "KmsKeyId": "string"
      },
      "StorageType": "string",
      "TtlDuration": {
         "Unit": "string",
         "Value": number
      }
   },
   "RecordIdentifierFeatureName": "string",
   "RoleArn": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "ThroughputConfig": {
      "ProvisionedReadCapacityUnits": number,
      "ProvisionedWriteCapacityUnits": number,
      "ThroughputMode": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Description

A free-form description of a `FeatureGroup`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Required: No

EventTimeFeatureName

The name of the feature that stores the `EventTime` of a `Record` in a `FeatureGroup`.

An `EventTime` is a point in time when a new event occurs that corresponds to the creation or update of a `Record` in a `FeatureGroup`. All `Records` in the `FeatureGroup` must have a corresponding `EventTime`.

An `EventTime` can be a `String` or `Fractional`.

- `Fractional`: `EventTime` feature values must be a Unix timestamp in seconds.
- `String`: `EventTime` feature values must be an ISO-8601 string in the format. The following formats are supported `yyyy-MM-dd'T'HH:mm:ssZ` and `yyyy-MM-dd'T'HH:mm:ss.SSSZ` where `yyyy`, `MM`, and `dd` represent the year, month, and day respectively and `HH`, `mm`, `ss`, and if applicable, `SSS` represent the hour, month, second and milliseconds respsectively. `'T'` and `Z` are constants.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

Required: Yes

FeatureDefinitions

A list of `Feature` names and types. `Name` and `Type` is compulsory per `Feature`.

Valid feature `FeatureTypes` are `Integral`, `Fractional` and `String`.

`FeatureNames` cannot be any of the following: `is_deleted`, `write_time`, `api_invocation_time`

You can create up to 2,500 `FeatureDefinitions` per `FeatureGroup`.

Type: Array of FeatureDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 2500 items.

Required: Yes

FeatureGroupName

The name of the `FeatureGroup`. The name must be unique within an AWS Region in an AWS account.

The name:

- Must start with an alphanumeric character.
- Can only include alphanumeric characters, underscores, and hyphens. Spaces are not allowed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63}`

Required: Yes

OfflineStoreConfig

Use this to configure an `OfflineFeatureStore`. This parameter allows you to specify:

- The Amazon Simple Storage Service (Amazon S3) location of an `OfflineStore`.
- A configuration for an AWS Glue or AWS Hive data catalog.
- An KMS encryption key to encrypt the Amazon S3 location used for `OfflineStore`. If KMS encryption key is not specified, by default we encrypt all data at rest using AWS KMS key. By defining your bucket-level key for SSE, you can reduce AWS KMS requests costs by up to 99 percent.
- Format for the offline store table. Supported formats are Glue (Default) and Apache Iceberg.

To learn more about this parameter, see OfflineStoreConfig.

Type: OfflineStoreConfig object

Required: No

OnlineStoreConfig

You can turn the `OnlineStore` on or off by specifying `True` for the `EnableOnlineStore` flag in `OnlineStoreConfig`.

You can also include an AWS KMS key ID (`KMSKeyId`) for at-rest encryption of the `OnlineStore`.

The default value is `False`.

Type: OnlineStoreConfig object

Required: No

RecordIdentifierFeatureName

The name of the `Feature` whose value uniquely identifies a `Record` defined in the `FeatureStore`. Only the latest record per identifier value will be stored in the `OnlineStore`. `RecordIdentifierFeatureName` must be one of feature definitions' names.

You use the `RecordIdentifierFeatureName` to access data in a `FeatureStore`.

This name:

- Must start with an alphanumeric character.
- Can only contains alphanumeric characters, hyphens, underscores. Spaces are not allowed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of the IAM execution role used to persist data into the `OfflineStore` if an `OfflineStoreConfig` is provided.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

Tags

Tags used to identify `Features` in each `FeatureGroup`.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

ThroughputConfig

Used to set feature group throughput configuration. There are two modes: `ON_DEMAND` and `PROVISIONED`. With on-demand mode, you are charged for data reads and writes that your application performs on your feature group. You do not need to specify read and write throughput because Feature Store accommodates your workloads as they ramp up and down. You can switch a feature group to on-demand only once in a 24 hour period. With provisioned throughput mode, you specify the read and write capacity per second that you expect your application to require, and you are billed based on those limits. Exceeding provisioned throughput will result in your requests being throttled.

Note: `PROVISIONED` throughput mode is supported only for feature groups that are offline-only, or use the `Standard` tier online store.

Type: ThroughputConfig object

Required: No

Response Syntax

```
{
   "FeatureGroupArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

FeatureGroupArn

The Amazon Resource Name (ARN) of the `FeatureGroup`. This is a unique identifier for the feature group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateFlowDefinition

Service: Amazon SageMaker Service

Creates a flow definition.

Request Syntax

```
{
   "FlowDefinitionName": "string",
   "HumanLoopActivationConfig": {
      "HumanLoopActivationConditionsConfig": {
         "HumanLoopActivationConditions": "string"
      }
   },
   "HumanLoopConfig": {
      "HumanTaskUiArn": "string",
      "PublicWorkforceTaskPrice": {
         "AmountInUsd": {
            "Cents": number,
            "Dollars": number,
            "TenthFractionsOfACent": number
         }
      },
      "TaskAvailabilityLifetimeInSeconds": number,
      "TaskCount": number,
      "TaskDescription": "string",
      "TaskKeywords": [ "string" ],
      "TaskTimeLimitInSeconds": number,
      "TaskTitle": "string",
      "WorkteamArn": "string"
   },
   "HumanLoopRequestSource": {
      "AwsManagedHumanLoopRequestSource": "string"
   },
   "OutputConfig": {
      "KmsKeyId": "string",
      "S3OutputPath": "string"
   },
   "RoleArn": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
```

```
        }
    ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

FlowDefinitionName

The name of your flow definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9]){0,62}`

Required: Yes

HumanLoopActivationConfig

An object containing information about the events that trigger a human workflow.

Type: HumanLoopActivationConfig object

Required: No

HumanLoopConfig

An object containing information about the tasks the human reviewers will perform.

Type: HumanLoopConfig object

Required: No

HumanLoopRequestSource

Container for configuring the source of human task requests. Use to specify if Amazon Rekognition or Amazon Textract is used as an integration source.

Type: HumanLoopRequestSource object

Required: No

OutputConfig

An object containing information about where the human review results will be uploaded.

Type: FlowDefinitionOutputConfig object

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of the role needed to call other services on your behalf. For example, `arn:aws:iam::1234567890:role/service-role/AmazonSageMaker-ExecutionRole-20180111T151298`.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Tags

An array of key-value pairs that contain metadata to help you categorize and organize a flow definition. Each tag consists of a key and a value, both of which you define.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "FlowDefinitionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

FlowDefinitionArn

The Amazon Resource Name (ARN) of the flow definition you create.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]+:[0-9]{12}:flow-definition/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateHub

Service: Amazon SageMaker Service

Create a hub.

Request Syntax

```
{
   "HubDescription": "string",
   "HubDisplayName": "string",
   "HubName": "string",
   "HubSearchKeywords": [ "string" ],
   "S3StorageConfig": {
      "S3OutputPath": "string"
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HubDescription

A description of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Pattern: `.*`

Required: Yes

HubDisplayName

The display name of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

HubName

The name of the hub to create.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubSearchKeywords

The searchable keywords for the hub.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[^A-Z]*`

Required: No

S3StorageConfig

The Amazon S3 storage configuration for the hub.

Type: HubS3StorageConfig object

Required: No

Tags

Any tags to associate with the hub.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "HubArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HubArn

The Amazon Resource Name (ARN) of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateHubContentPresignedUrls

Service: Amazon SageMaker Service

Creates presigned URLs for accessing hub content artifacts. This operation generates time-limited, secure URLs that allow direct download of model artifacts and associated files from Amazon SageMaker hub content, including gated models that require end-user license agreement acceptance.

Request Syntax

```
{
   "AccessConfig": {
      "AcceptEula": boolean,
      "ExpectedS3Url": "string"
   },
   "HubContentName": "string",
   "HubContentType": "string",
   "HubContentVersion": "string",
   "HubName": "string",
   "MaxResults": number,
   "NextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AccessConfig

Configuration settings for accessing the hub content, including end-user license agreement acceptance for gated models and expected S3 URL validation.

Type: PresignedUrlAccessConfig object

Required: No

HubContentName

The name of the hub content for which to generate presigned URLs. This identifies the specific model or content within the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubContentType

The type of hub content to access. Valid values include `Model`, `Notebook`, and `ModelReference`.

Type: String

Valid Values: `Model | Notebook | ModelReference | DataSet | JsonDoc`

Required: Yes

HubContentVersion

The version of the hub content. If not specified, the latest version is used.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: No

HubName

The name or Amazon Resource Name (ARN) of the hub that contains the content. For public content, use `SageMakerPublicHub`.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

MaxResults

The maximum number of presigned URLs to return in the response. Default value is 100. Large models may contain hundreds of files, requiring pagination to retrieve all URLs.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

A token for pagination. Use this token to retrieve the next set of presigned URLs when the response is truncated.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "AuthorizedUrlConfigs": [
      {
         "LocalPath": "string",
         "Url": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AuthorizedUrlConfigs

An array of authorized URL configurations, each containing a presigned URL and its corresponding local file path for proper file organization during download.

Type: Array of AuthorizedUrl objects

NextToken

A token for pagination. If present, indicates that more presigned URLs are available. Use this token in a subsequent request to retrieve additional URLs.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateHubContentReference

Service: Amazon SageMaker Service

Create a hub content reference in order to add a model in the JumpStart public hub to a private hub.

Request Syntax

```
{
   "HubContentName": "string",
   "HubName": "string",
   "MinVersion": "string",
   "SageMakerPublicHubContentArn": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HubContentName

The name of the hub content to reference.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

HubName

The name of the hub to add the hub content reference to.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

MinVersion

The minimum version of the hub content to reference.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: No

SageMakerPublicHubContentArn

The ARN of the public hub content to reference.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:aws:hub-content\/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}\/Model\/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: Yes

Tags

Any tags associated with the hub content to reference.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
```

```
   "HubArn": "string",
   "HubContentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HubArn

The ARN of the hub that the hub content reference was added to.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

HubContentArn

The ARN of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateHumanTaskUi

Service: Amazon SageMaker Service

Defines the settings you will use for the human review workflow user interface. Reviewers will see a three-panel interface with an instruction area, the item to review, and an input area.

Request Syntax

```
{
   "HumanTaskUiName": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "UiTemplate": {
      "Content": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HumanTaskUiName

The name of the user interface you are creating.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9])*`

Required: Yes

Tags

An array of key-value pairs that contain metadata to help you categorize and organize a human review workflow user interface. Each tag consists of a key and a value, both of which you define.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

UiTemplate

The Liquid template for the worker user interface.

Type: UiTemplate object

Required: Yes

Response Syntax

```
{
   "HumanTaskUiArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HumanTaskUiArn

The Amazon Resource Name (ARN) of the human review workflow user interface you create.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]+:[0-9]{12}:human-task-ui/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateHyperParameterTuningJob

Service: Amazon SageMaker Service

Starts a hyperparameter tuning job. A hyperparameter tuning job finds the best version of a model by running many training jobs on your dataset using the algorithm you choose and values for hyperparameters within ranges that you specify. It then chooses the hyperparameter values that result in a model that performs the best, as measured by an objective metric that you choose.

A hyperparameter tuning job automatically creates Amazon SageMaker experiments, trials, and trial components for each training job that it runs. You can view these entities in Amazon SageMaker Studio. For more information, see View Experiments, Trials, and Trial Components.

> **Important**
>
> Do not include any security-sensitive information including account access IDs, secrets, or tokens in any hyperparameter fields. As part of the shared responsibility model, you are responsible for any potential exposure, unauthorized access, or compromise of your sensitive data if caused by any security-sensitive information included in the request hyperparameter variable or plain text fields..

Request Syntax

```
{
   "Autotune": {
      "Mode": "string"
   },
   "HyperParameterTuningJobConfig": {
      "HyperParameterTuningJobObjective": {
         "MetricName": "string",
         "Type": "string"
      },
      "ParameterRanges": {
         "AutoParameters": [
            {
               "Name": "string",
               "ValueHint": "string"
            }
         ],
         "CategoricalParameterRanges": [
            {
```

```
            "Name": "string",
            "Values": [ "string" ]
         }
      ],
      "ContinuousParameterRanges": [
         {
            "MaxValue": "string",
            "MinValue": "string",
            "Name": "string",
            "ScalingType": "string"
         }
      ],
      "IntegerParameterRanges": [
         {
            "MaxValue": "string",
            "MinValue": "string",
            "Name": "string",
            "ScalingType": "string"
         }
      ]
   },
   "RandomSeed": number,
   "ResourceLimits": {
      "MaxNumberOfTrainingJobs": number,
      "MaxParallelTrainingJobs": number,
      "MaxRuntimeInSeconds": number
   },
   "Strategy": "string",
   "StrategyConfig": {
      "HyperbandStrategyConfig": {
         "MaxResource": number,
         "MinResource": number
      }
   },
   "TrainingJobEarlyStoppingType": "string",
   "TuningJobCompletionCriteria": {
      "BestObjectiveNotImproving": {
         "MaxNumberOfTrainingJobsNotImproving": number
      },
      "ConvergenceDetected": {
         "CompleteOnConvergence": "string"
      },
      "TargetObjectiveMetricValue": number
   }
```

```
    },
    "HyperParameterTuningJobName": "string",
    "Tags": [
       {
          "Key": "string",
          "Value": "string"
       }
    ],
    "TrainingJobDefinition": {
       "AlgorithmSpecification": {
          "AlgorithmName": "string",
          "MetricDefinitions": [
             {
                "Name": "string",
                "Regex": "string"
             }
          ],
          "TrainingImage": "string",
          "TrainingInputMode": "string"
       },
       "CheckpointConfig": {
          "LocalPath": "string",
          "S3Uri": "string"
       },
       "DefinitionName": "string",
       "EnableInterContainerTrafficEncryption": boolean,
       "EnableManagedSpotTraining": boolean,
       "EnableNetworkIsolation": boolean,
       "Environment": {
          "string" : "string"
       },
       "HyperParameterRanges": {
          "AutoParameters": [
             {
                "Name": "string",
                "ValueHint": "string"
             }
          ],
          "CategoricalParameterRanges": [
             {
                "Name": "string",
                "Values": [ "string" ]
             }
          ],
```

```
            "ContinuousParameterRanges": [
               {
                  "MaxValue": "string",
                  "MinValue": "string",
                  "Name": "string",
                  "ScalingType": "string"
               }
            ],
            "IntegerParameterRanges": [
               {
                  "MaxValue": "string",
                  "MinValue": "string",
                  "Name": "string",
                  "ScalingType": "string"
               }
            ]
         },
         "HyperParameterTuningResourceConfig": {
            "AllocationStrategy": "string",
            "InstanceConfigs": [
               {
                  "InstanceCount": number,
                  "InstanceType": "string",
                  "VolumeSizeInGB": number
               }
            ],
            "InstanceCount": number,
            "InstanceType": "string",
            "VolumeKmsKeyId": "string",
            "VolumeSizeInGB": number
         },
         "InputDataConfig": [
            {
               "ChannelName": "string",
               "CompressionType": "string",
               "ContentType": "string",
               "DataSource": {
                  "DatasetSource": {
                     "DatasetArn": "string"
                  },
                  "FileSystemDataSource": {
                     "DirectoryPath": "string",
                     "FileSystemAccessMode": "string",
                     "FileSystemId": "string",
```

```
                "FileSystemType": "string"
             },
             "S3DataSource": {
                "AttributeNames": [ "string" ],
                "HubAccessConfig": {
                   "HubContentArn": "string"
                },
                "InstanceGroupNames": [ "string" ],
                "ModelAccessConfig": {
                   "AcceptEula": boolean
                },
                "S3DataDistributionType": "string",
                "S3DataType": "string",
                "S3Uri": "string"
             }
          },
          "InputMode": "string",
          "RecordWrapperType": "string",
          "ShuffleConfig": {
             "Seed": number
          }
       }
    ],
    "OutputDataConfig": {
       "CompressionType": "string",
       "KmsKeyId": "string",
       "S3OutputPath": "string"
    },
    "ResourceConfig": {
       "InstanceCount": number,
       "InstanceGroups": [
          {
             "InstanceCount": number,
             "InstanceGroupName": "string",
             "InstanceType": "string"
          }
       ],
       "InstancePlacementConfig": {
          "EnableMultipleJobs": boolean,
          "PlacementSpecifications": [
             {
                "InstanceCount": number,
                "UltraServerId": "string"
             }
```

```
            ]
         },
         "InstanceType": "string",
         "KeepAlivePeriodInSeconds": number,
         "TrainingPlanArn": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
      },
      "RetryStrategy": {
         "MaximumRetryAttempts": number
      },
      "RoleArn": "string",
      "StaticHyperParameters": {
         "string" : "string"
      },
      "StoppingCondition": {
         "MaxPendingTimeInSeconds": number,
         "MaxRuntimeInSeconds": number,
         "MaxWaitTimeInSeconds": number
      },
      "TuningObjective": {
         "MetricName": "string",
         "Type": "string"
      },
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
   },
   "TrainingJobDefinitions": [
      {
         "AlgorithmSpecification": {
            "AlgorithmName": "string",
            "MetricDefinitions": [
               {
                  "Name": "string",
                  "Regex": "string"
               }
            ],
            "TrainingImage": "string",
            "TrainingInputMode": "string"
         },
         "CheckpointConfig": {
            "LocalPath": "string",
```

```
            "S3Uri": "string"
         },
         "DefinitionName": "string",
         "EnableInterContainerTrafficEncryption": boolean,
         "EnableManagedSpotTraining": boolean,
         "EnableNetworkIsolation": boolean,
         "Environment": {
            "string" : "string"
         },
         "HyperParameterRanges": {
            "AutoParameters": [
               {
                  "Name": "string",
                  "ValueHint": "string"
               }
            ],
            "CategoricalParameterRanges": [
               {
                  "Name": "string",
                  "Values": [ "string" ]
               }
            ],
            "ContinuousParameterRanges": [
               {
                  "MaxValue": "string",
                  "MinValue": "string",
                  "Name": "string",
                  "ScalingType": "string"
               }
            ],
            "IntegerParameterRanges": [
               {
                  "MaxValue": "string",
                  "MinValue": "string",
                  "Name": "string",
                  "ScalingType": "string"
               }
            ]
         },
         "HyperParameterTuningResourceConfig": {
            "AllocationStrategy": "string",
            "InstanceConfigs": [
               {
                  "InstanceCount": number,
```

```
                "InstanceType": "string",
                "VolumeSizeInGB": number
             }
          ],
          "InstanceCount": number,
          "InstanceType": "string",
          "VolumeKmsKeyId": "string",
          "VolumeSizeInGB": number
       },
       "InputDataConfig": [
          {
             "ChannelName": "string",
             "CompressionType": "string",
             "ContentType": "string",
             "DataSource": {
                "DatasetSource": {
                   "DatasetArn": "string"
                },
                "FileSystemDataSource": {
                   "DirectoryPath": "string",
                   "FileSystemAccessMode": "string",
                   "FileSystemId": "string",
                   "FileSystemType": "string"
                },
                "S3DataSource": {
                   "AttributeNames": [ "string" ],
                   "HubAccessConfig": {
                      "HubContentArn": "string"
                   },
                   "InstanceGroupNames": [ "string" ],
                   "ModelAccessConfig": {
                      "AcceptEula": boolean
                   },
                   "S3DataDistributionType": "string",
                   "S3DataType": "string",
                   "S3Uri": "string"
                }
             },
             "InputMode": "string",
             "RecordWrapperType": "string",
             "ShuffleConfig": {
                "Seed": number
             }
          }
```

```
        ],
        "OutputDataConfig": {
           "CompressionType": "string",
           "KmsKeyId": "string",
           "S3OutputPath": "string"
        },
        "ResourceConfig": {
           "InstanceCount": number,
           "InstanceGroups": [
              {
                 "InstanceCount": number,
                 "InstanceGroupName": "string",
                 "InstanceType": "string"
              }
           ],
           "InstancePlacementConfig": {
              "EnableMultipleJobs": boolean,
              "PlacementSpecifications": [
                 {
                    "InstanceCount": number,
                    "UltraServerId": "string"
                 }
              ]
           },
           "InstanceType": "string",
           "KeepAlivePeriodInSeconds": number,
           "TrainingPlanArn": "string",
           "VolumeKmsKeyId": "string",
           "VolumeSizeInGB": number
        },
        "RetryStrategy": {
           "MaximumRetryAttempts": number
        },
        "RoleArn": "string",
        "StaticHyperParameters": {
           "string" : "string"
        },
        "StoppingCondition": {
           "MaxPendingTimeInSeconds": number,
           "MaxRuntimeInSeconds": number,
           "MaxWaitTimeInSeconds": number
        },
        "TuningObjective": {
           "MetricName": "string",
```

```
            "Type": "string"
         },
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      }
   ],
   "WarmStartConfig": {
      "ParentHyperParameterTuningJobs": [
         {
            "HyperParameterTuningJobName": "string"
         }
      ],
      "WarmStartType": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Autotune

Configures SageMaker Automatic model tuning (AMT) to automatically find optimal parameters for the following fields:

- ParameterRanges: The names and ranges of parameters that a hyperparameter tuning job can optimize.
- ResourceLimits: The maximum resources that can be used for a training job. These resources include the maximum number of training jobs, the maximum runtime of a tuning job, and the maximum number of training jobs to run at the same time.
- TrainingJobEarlyStoppingType: A flag that specifies whether or not to use early stopping for training jobs launched by a hyperparameter tuning job.
- RetryStrategy: The number of times to retry a training job.
- Strategy: Specifies how hyperparameter tuning chooses the combinations of hyperparameter values to use for the training jobs that it launches.
- ConvergenceDetected: A flag to indicate that Automatic model tuning (AMT) has detected model convergence.

Type: Autotune object

Required: No

HyperParameterTuningJobConfig

The HyperParameterTuningJobConfig object that describes the tuning job, including the search strategy, the objective metric used to evaluate training jobs, ranges of parameters to search, and resource limits for the tuning job. For more information, see How Hyperparameter Tuning Works.

Type: HyperParameterTuningJobConfig object

Required: Yes

HyperParameterTuningJobName

The name of the tuning job. This name is the prefix for the names of all training jobs that this tuning job launches. The name must be unique within the same AWS account and AWS Region. The name must have 1 to 32 characters. Valid characters are a-z, A-Z, 0-9, and : + = @ _ % - (hyphen). The name is not case sensitive.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Tags that you specify for the tuning job are also added to all training jobs that the tuning job launches.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TrainingJobDefinition

The HyperParameterTrainingJobDefinition object that describes the training jobs that this tuning job launches, including static hyperparameters, input data configuration, output data configuration, resource configuration, and stopping condition.

Type: HyperParameterTrainingJobDefinition object

Required: No

TrainingJobDefinitions

A list of the HyperParameterTrainingJobDefinition objects launched for this tuning job.

Type: Array of HyperParameterTrainingJobDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

WarmStartConfig

Specifies the configuration for starting the hyperparameter tuning job using one or more previous tuning jobs as a starting point. The results of previous tuning jobs are used to inform which combinations of hyperparameters to search over in the new tuning job.

All training jobs launched by the new hyperparameter tuning job are evaluated by using the objective metric. If you specify `IDENTICAL_DATA_AND_ALGORITHM` as the `WarmStartType` value for the warm start configuration, the training job that performs the best in the new tuning job is compared to the best training jobs from the parent tuning jobs. From these, the training job that performs the best as measured by the objective metric is returned as the overall best training job.

> **Note**
>
> All training jobs launched by parent hyperparameter tuning jobs and the new hyperparameter tuning jobs count against the limit of training jobs for the tuning job.

Type: HyperParameterTuningJobWarmStartConfig object

Required: No

Response Syntax

```
{
   "HyperParameterTuningJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HyperParameterTuningJobArn

The Amazon Resource Name (ARN) of the tuning job. SageMaker assigns an ARN to a hyperparameter tuning job when you create it.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:hyper-parameter-tuning-job/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateImage

Service: Amazon SageMaker Service

Creates a custom SageMaker AI image. A SageMaker AI image is a set of image versions. Each image version represents a container image stored in Amazon ECR. For more information, see Bring your own SageMaker AI image.

Request Syntax

```
{
   "Description": "string",
   "DisplayName": "string",
   "ImageName": "string",
   "RoleArn": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Description

The description of the image.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 512.

Pattern: `.*`

Required: No

DisplayName

The display name of the image. If not provided, `ImageName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `\S(.*\S)?`

Required: No

ImageName

The name of the image. Must be unique to your account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

RoleArn

The ARN of an IAM role that enables Amazon SageMaker AI to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Tags

A list of tags to apply to the image.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
```

```
   "ImageArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ImageArn

The ARN of the image.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image/[a-zA-Z0-9]([-.]?[a-zA-Z0-9])*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateImageVersion

Service: Amazon SageMaker Service

Creates a version of the SageMaker AI image specified by `ImageName`. The version represents the Amazon ECR container image specified by `BaseImage`.

Request Syntax

```
{
   "Aliases": [ "string" ],
   "BaseImage": "string",
   "ClientToken": "string",
   "Horovod": boolean,
   "ImageName": "string",
   "JobType": "string",
   "MLFramework": "string",
   "Processor": "string",
   "ProgrammingLang": "string",
   "ReleaseNotes": "string",
   "VendorGuidance": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Aliases

A list of aliases created with the image version.

Type: Array of strings

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(?!^[.-])^([a-zA-Z0-9-_.]+)`

Required: No

BaseImage

The registry path of the container image to use as the starting point for this version. The path is an Amazon ECR URI in the following format:

`<acct-id>.dkr.ecr.<region>.amazonaws.com/<repo-name[:tag] or [@digest]>`

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.*`

Required: Yes

ClientToken

A unique ID. If not specified, the AWS CLI and AWS SDKs, such as the SDK for Python (Boto3), add a unique value to the call.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 36.

Pattern: `[a-zA-Z0-9-]+`

Required: Yes

Horovod

Indicates Horovod compatibility.

Type: Boolean

Required: No

ImageName

The `ImageName` of the `Image` to create a version of.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

JobType

Indicates SageMaker AI job type compatibility.

- `TRAINING`: The image version is compatible with SageMaker AI training jobs.
- `INFERENCE`: The image version is compatible with SageMaker AI inference jobs.
- `NOTEBOOK_KERNEL`: The image version is compatible with SageMaker AI notebook kernels.

Type: String

Valid Values: `TRAINING | INFERENCE | NOTEBOOK_KERNEL`

Required: No

MLFramework

The machine learning framework vended in the image version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z]+ ?\d+\.\d+(\.\d+)?`

Required: No

Processor

Indicates CPU or GPU compatibility.

- `CPU`: The image version is compatible with CPU.
- `GPU`: The image version is compatible with GPU.

Type: String

Valid Values: `CPU | GPU`

Required: No

ProgrammingLang

The supported programming language and its version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z]+ ?\d+\.\d+(\.\d+)?`

Required: No

ReleaseNotes

The maintainer description of the image version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.*`

Required: No

VendorGuidance

The stability of the image version, specified by the maintainer.

- `NOT_PROVIDED`: The maintainers did not provide a status for image version stability.
- `STABLE`: The image version is stable.
- `TO_BE_ARCHIVED`: The image version is set to be archived. Custom image versions that are set to be archived are automatically archived after three months.
- `ARCHIVED`: The image version is archived. Archived image versions are not searchable and are no longer actively supported.

Type: String

Valid Values: `NOT_PROVIDED | STABLE | TO_BE_ARCHIVED | ARCHIVED`

Required: No

Response Syntax

```
{
   "ImageVersionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ImageVersionArn

The ARN of the image version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image-version/[a-z0-9]([-.]?[a-z0-9])*/[0-9]+|None)`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateInferenceComponent

Service: Amazon SageMaker Service

Creates an inference component, which is a SageMaker AI hosting object that you can use to deploy a model to an endpoint. In the inference component settings, you specify the model, the endpoint, and how the model utilizes the resources that the endpoint hosts. You can optimize resource utilization by tailoring how the required CPU cores, accelerators, and memory are allocated. You can deploy multiple inference components to an endpoint, where each inference component contains one model and the resource utilization needs for that individual model. After you deploy an inference component, you can directly invoke the associated model when you use the InvokeEndpoint API action.

Request Syntax

```
{
   "EndpointName": "string",
   "InferenceComponentName": "string",
   "RuntimeConfig": {
      "CopyCount": number
   },
   "Specification": {
      "BaseInferenceComponentName": "string",
      "ComputeResourceRequirements": {
         "MaxMemoryRequiredInMb": number,
         "MinMemoryRequiredInMb": number,
         "NumberOfAcceleratorDevicesRequired": number,
         "NumberOfCpuCoresRequired": number
      },
      "Container": {
         "ArtifactUrl": "string",
         "ContainerMetricsConfig": {
            "MetricsEndpoints": [
               {
                  "MetricPublishFrequencyInSeconds": number,
                  "MetricsEndpointPath": "string"
               }
            ]
         },
         "Environment": {
            "string" : "string"
         },
         "Image": "string"
```

```
      },
      "DataCacheConfig": {
         "EnableCaching": boolean
      },
      "InstanceType": "string",
      "ModelName": "string",
      "SchedulingConfig": {
         "AvailabilityZoneBalance": {
            "EnforcementMode": "string",
            "MaxImbalance": number
         },
         "PlacementStrategy": "string"
      },
      "StartupParameters": {
         "ContainerStartupHealthCheckTimeoutInSeconds": number,
         "ModelDataDownloadTimeoutInSeconds": number
      }
   },
   "Specifications": [
      {
         "BaseInferenceComponentName": "string",
         "ComputeResourceRequirements": {
            "MaxMemoryRequiredInMb": number,
            "MinMemoryRequiredInMb": number,
            "NumberOfAcceleratorDevicesRequired": number,
            "NumberOfCpuCoresRequired": number
         },
         "Container": {
            "ArtifactUrl": "string",
            "ContainerMetricsConfig": {
               "MetricsEndpoints": [
                  {
                     "MetricPublishFrequencyInSeconds": number,
                     "MetricsEndpointPath": "string"
                  }
               ]
            },
            "Environment": {
               "string" : "string"
            },
            "Image": "string"
         },
         "DataCacheConfig": {
            "EnableCaching": boolean
```

```
         },
         "InstanceType": "string",
         "ModelName": "string",
         "SchedulingConfig": {
            "AvailabilityZoneBalance": {
               "EnforcementMode": "string",
               "MaxImbalance": number
            },
            "PlacementStrategy": "string"
         },
         "StartupParameters": {
            "ContainerStartupHealthCheckTimeoutInSeconds": number,
            "ModelDataDownloadTimeoutInSeconds": number
         }
      }
   ],
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "VariantName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EndpointName

The name of an existing endpoint where you host the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

InferenceComponentName

A unique name to assign to the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

RuntimeConfig

Runtime settings for a model that is deployed with an inference component.

Type: InferenceComponentRuntimeConfig object

Required: No

Specification

Details about the resources to deploy with this inference component, including the model, container, and compute resources.

Type: InferenceComponentSpecification object

Required: No

Specifications

A list of specification objects for the inference component, one per instance type. Use this parameter when you want to deploy a different model or resource configuration for the inference component on each instance type. You can use either this parameter or the singular `Specification` parameter, but not both.

Type: Array of InferenceComponentSpecification objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

Tags

A list of key-value pairs associated with the model. For more information, see Tagging AWS resources in the *AWS General Reference*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

VariantName

The name of an existing production variant where you host the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Response Syntax

```
{
   "InferenceComponentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

InferenceComponentArn

The Amazon Resource Name (ARN) of the inference component.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateInferenceExperiment

Service: Amazon SageMaker Service

Creates an inference experiment using the configurations specified in the request.

Use this API to setup and schedule an experiment to compare model variants on a Amazon SageMaker inference endpoint. For more information about inference experiments, see Shadow tests.

Amazon SageMaker begins your experiment at the scheduled time and routes traffic to your endpoint's model variants based on your specified configuration.

While the experiment is in progress or after it has concluded, you can view metrics that compare your model variants. For more information, see View, monitor, and edit shadow tests.

Request Syntax

```
{
   "DataStorageConfig": {
      "ContentType": {
         "CsvContentTypes": [ "string" ],
         "JsonContentTypes": [ "string" ]
      },
      "Destination": "string",
      "KmsKey": "string"
   },
   "Description": "string",
   "EndpointName": "string",
   "KmsKey": "string",
   "ModelVariants": [
      {
         "InfrastructureConfig": {
            "InfrastructureType": "string",
            "RealTimeInferenceConfig": {
               "InstanceCount": number,
               "InstanceType": "string"
            }
         },
         "ModelName": "string",
         "VariantName": "string"
      }
   ],
```

```
   "Name": "string",
   "RoleArn": "string",
   "Schedule": {
      "EndTime": number,
      "StartTime": number
   },
   "ShadowModeConfig": {
      "ShadowModelVariants": [
         {
            "SamplingPercentage": number,
            "ShadowModelVariantName": "string"
         }
      ],
      "SourceModelVariantName": "string"
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "Type": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DataStorageConfig

The Amazon S3 location and configuration for storing inference request and response data.

This is an optional parameter that you can use for data capture. For more information, see Capture data.

Type: InferenceExperimentDataStorageConfig object

Required: No

Description

A description for the inference experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

EndpointName

The name of the Amazon SageMaker endpoint on which you want to run the inference experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

KmsKey

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt data on the storage volume attached to the ML compute instance that hosts the endpoint. The `KmsKey` can be any of the following formats:

- KMS key ID

  `"1234abcd-12ab-34cd-56ef-1234567890ab"`

- Amazon Resource Name (ARN) of a KMS key

  `"arn:aws:kms:us-west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab"`

- KMS key Alias

  `"alias/ExampleAlias"`

- Amazon Resource Name (ARN) of a KMS key Alias

  `"arn:aws:kms:us-west-2:111122223333:alias/ExampleAlias"`

If you use a KMS key ID or an alias of your KMS key, the Amazon SageMaker execution role must include permissions to call `kms:Encrypt`. If you don't provide a KMS key ID, Amazon

SageMaker uses the default KMS key for Amazon S3 for your role's account. Amazon SageMaker uses server-side encryption with KMS managed keys for `OutputDataConfig`. If you use a bucket policy with an `s3:PutObject` permission that only allows objects with server-side encryption, set the condition key of `s3:x-amz-server-side-encryption` to `"aws:kms"`. For more information, see KMS managed Encryption Keys in the *Amazon Simple Storage Service Developer Guide.*

The KMS key policy must grant permission to the IAM role that you specify in your `CreateEndpoint` and `UpdateEndpoint` requests. For more information, see Using Key Policies in AWS KMS in the *AWS Key Management Service Developer Guide*.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

ModelVariants

An array of `ModelVariantConfig` objects. There is one for each variant in the inference experiment. Each `ModelVariantConfig` object in the array describes the infrastructure configuration for the corresponding variant.

Type: Array of ModelVariantConfig objects

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Required: Yes

Name

The name for the inference experiment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

RoleArn

The ARN of the IAM role that Amazon SageMaker can assume to access model artifacts and container images, and manage Amazon SageMaker Inference endpoints for model deployment.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Schedule

The duration for which you want the inference experiment to run. If you don't specify this field, the experiment automatically starts immediately upon creation and concludes after 7 days.

Type: InferenceExperimentSchedule object

Required: No

ShadowModeConfig

The configuration of `ShadowMode` inference experiment type. Use this field to specify a production variant which takes all the inference requests, and a shadow variant to which Amazon SageMaker replicates a percentage of the inference requests. For the shadow variant also specify the percentage of requests that Amazon SageMaker replicates.

Type: ShadowModeConfig object

Required: Yes

Tags

Array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging your AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Type

The type of the inference experiment that you want to run. The following types of experiments are possible:

- ShadowMode: You can use this type to validate a shadow variant. For more information, see Shadow tests.

Type: String

Valid Values: ShadowMode

Required: Yes

Response Syntax

```
{
   "InferenceExperimentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

InferenceExperimentArn

The ARN for your inference experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:inference-experiment/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateInferenceRecommendationsJob

Service: Amazon SageMaker Service

Starts a recommendation job. You can create either an instance recommendation or load test job.

Request Syntax

```
{
   "InputConfig": {
      "ContainerConfig": {
         "DataInputConfig": "string",
         "Domain": "string",
         "Framework": "string",
         "FrameworkVersion": "string",
         "NearestModelName": "string",
         "PayloadConfig": {
            "SamplePayloadUrl": "string",
            "SupportedContentTypes": [ "string" ]
         },
         "SupportedEndpointType": "string",
         "SupportedInstanceTypes": [ "string" ],
         "SupportedResponseMIMETypes": [ "string" ],
         "Task": "string"
      },
      "EndpointConfigurations": [
         {
            "EnvironmentParameterRanges": {
               "CategoricalParameterRanges": [
                  {
                     "Name": "string",
                     "Value": [ "string" ]
                  }
               ]
            },
            "InferenceSpecificationName": "string",
            "InstanceType": "string",
            "ServerlessConfig": {
               "MaxConcurrency": number,
               "MemorySizeInMB": number,
               "ProvisionedConcurrency": number
            }
         }
      ],
```

```
      "Endpoints": [
         {
            "EndpointName": "string"
         }
      ],
      "JobDurationInSeconds": number,
      "ModelName": "string",
      "ModelPackageVersionArn": "string",
      "ResourceLimit": {
         "MaxNumberOfTests": number,
         "MaxParallelOfTests": number
      },
      "TrafficPattern": {
         "Phases": [
            {
               "DurationInSeconds": number,
               "InitialNumberOfUsers": number,
               "SpawnRate": number
            }
         ],
         "Stairs": {
            "DurationInSeconds": number,
            "NumberOfSteps": number,
            "UsersPerStep": number
         },
         "TrafficType": "string"
      },
      "VolumeKmsKeyId": "string",
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
   },
   "JobDescription": "string",
   "JobName": "string",
   "JobType": "string",
   "OutputConfig": {
      "CompiledOutputConfig": {
         "S3OutputUri": "string"
      },
      "KmsKeyId": "string"
   },
   "RoleArn": "string",
   "StoppingConditions": {
```

```
      "FlatInvocations": "string",
      "MaxInvocations": number,
      "ModelLatencyThresholds": [
         {
            "Percentile": "string",
            "ValueInMilliseconds": number
         }
      ]
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

InputConfig

Provides information about the versioned model package Amazon Resource Name (ARN), the traffic pattern, and endpoint configurations.

Type: RecommendationJobInputConfig object

Required: Yes

JobDescription

Description of the recommendation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Required: No

JobName

A name for the recommendation job. The name must be unique within the AWS Region and within your AWS account. The job name is passed down to the resources created by the

recommendation job. The names of resources (such as the model, endpoint configuration, endpoint, and compilation) that are prefixed with the job name are truncated at 40 characters.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: Yes

JobType

Defines the type of recommendation job. Specify `Default` to initiate an instance recommendation and `Advanced` to initiate a load test. If left unspecified, Amazon SageMaker Inference Recommender will run an instance recommendation (`DEFAULT`) job.

Type: String

Valid Values: `Default | Advanced`

Required: Yes

OutputConfig

Provides information about the output artifacts and the KMS key to use for Amazon S3 server-side encryption.

Type: RecommendationJobOutputConfig object

Required: No

RoleArn

The Amazon Resource Name (ARN) of an IAM role that enables Amazon SageMaker to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

StoppingConditions

A set of conditions for stopping a recommendation job. If any of the conditions are met, the job is automatically stopped.

Type: RecommendationJobStoppingConditions object

Required: No

Tags

The metadata that you apply to AWS resources to help you categorize and organize them. Each tag consists of a key and a value, both of which you define. For more information, see Tagging AWS Resources in the AWS General Reference.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "JobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobArn

The Amazon Resource Name (ARN) of the recommendation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:inference-recommendations-job/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateJob

Service: Amazon SageMaker Service

Creates a model customization job in Amazon SageMaker. A job runs a workload based on the job category and configuration you provide. You specify the job category, a schema-versioned configuration document, and an IAM role that grants Amazon SageMaker permission to access resources on your behalf.

Use the `AgentRFT` category to fine-tune a model using multi-turn reinforcement learning with reward signals. Use the `AgentRFTEvaluation` category to evaluate a fine-tuned or base model by running multi-turn rollouts against a held-out prompt dataset and computing metrics such as pass@k and mean reward.

Before creating a job, call `ListJobSchemaVersions` and `DescribeJobSchemaVersion` to retrieve the configuration schema for your job category. The `JobConfigDocument` must conform to the schema specified by `JobConfigSchemaVersion`.

The following operations are related to `CreateJob`:

- `DescribeJob`
- `ListJobs`
- `StopJob`
- `DeleteJob`
- `ListJobSchemaVersions`
- `DescribeJobSchemaVersion`

Request Syntax

```
{
   "JobCategory": "string",
   "JobConfigDocument": "string",
   "JobConfigSchemaVersion": "string",
   "JobName": "string",
   "RoleArn": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
```

```
        }
    ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobCategory

The category of the job. The category determines the type of workload that the job runs.

Type: String

Valid Values: `AgentRFT | AgentRFTEvaluation`

Required: Yes

JobConfigDocument

The JSON configuration document for the job. The document must conform to the schema specified by `JobConfigSchemaVersion`. Use `DescribeJobSchemaVersion` to retrieve the schema for validation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 262144.

Required: Yes

JobConfigSchemaVersion

The version of the configuration schema to use for the job configuration document. Use `ListJobSchemaVersions` to get available schema versions for a job category.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 16.

Pattern: `\d+\.\d+\.\d+`

Required: Yes

JobName

The name of the job. The name must be unique within your account and AWS Region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of the IAM role that Amazon SageMaker assumes to perform the job. The role must have the necessary permissions to access the resources required by the job configuration.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Tags

An array of key-value pairs to apply to the job as tags. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "JobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobArn

The Amazon Resource Name (ARN) of the job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:(aws[a-z\-]*):sagemaker:[a-z0-9\-]+:[0-9]{12}:job/[a-zA-Z0-9]+/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateLabelingJob

Service: Amazon SageMaker Service

Creates a job that uses workers to label the data objects in your input dataset. You can use the labeled data to train machine learning models.

You can select your workforce from one of three providers:

- A private workforce that you create. It can include employees, contractors, and outside experts. Use a private workforce when want the data to stay within your organization or when a specific set of skills is required.
- One or more vendors that you select from the AWS Marketplace. Vendors provide expertise in specific areas.
- The Amazon Mechanical Turk workforce. This is the largest workforce, but it should only be used for public data or data that has been stripped of any personally identifiable information.

You can also use *automated data labeling* to reduce the number of data objects that need to be labeled by a human. Automated data labeling uses *active learning* to determine if a data object can be labeled by machine or if it needs to be sent to a human worker. For more information, see Using Automated Data Labeling.

The data objects to be labeled are contained in an Amazon S3 bucket. You create a *manifest file* that describes the location of each object. For more information, see Using Input and Output Data.

The output can be used as the manifest file for another labeling job or as training data for your machine learning models.

You can use this operation to create a static labeling job or a streaming labeling job. A static labeling job stops if all data objects in the input manifest file identified in `ManifestS3Uri` have been labeled. A streaming labeling job runs perpetually until it is manually stopped, or remains idle for 10 days. You can send new data objects to an active (`InProgress`) streaming labeling job in real time. To learn how to create a static labeling job, see Create a Labeling Job (API) in the Amazon SageMaker Developer Guide. To learn how to create a streaming labeling job, see Create a Streaming Labeling Job.

Request Syntax

```
{
   "HumanTaskConfig": {
```

```
      "AnnotationConsolidationConfig": {
         "AnnotationConsolidationLambdaArn": "string"
      },
      "MaxConcurrentTaskCount": number,
      "NumberOfHumanWorkersPerDataObject": number,
      "PreHumanTaskLambdaArn": "string",
      "PublicWorkforceTaskPrice": {
         "AmountInUsd": {
            "Cents": number,
            "Dollars": number,
            "TenthFractionsOfACent": number
         }
      },
      "TaskAvailabilityLifetimeInSeconds": number,
      "TaskDescription": "string",
      "TaskKeywords": [ "string" ],
      "TaskTimeLimitInSeconds": number,
      "TaskTitle": "string",
      "UiConfig": {
         "HumanTaskUiArn": "string",
         "UiTemplateS3Uri": "string"
      },
      "WorkteamArn": "string"
   },
   "InputConfig": {
      "DataAttributes": {
         "ContentClassifiers": [ "string" ]
      },
      "DataSource": {
         "S3DataSource": {
            "ManifestS3Uri": "string"
         },
         "SnsDataSource": {
            "SnsTopicArn": "string"
         }
      }
   },
   "LabelAttributeName": "string",
   "LabelCategoryConfigS3Uri": "string",
   "LabelingJobAlgorithmsConfig": {
      "InitialActiveLearningModelArn": "string",
      "LabelingJobAlgorithmSpecificationArn": "string",
      "LabelingJobResourceConfig": {
         "VolumeKmsKeyId": "string",
```

```
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      }
   },
   "LabelingJobName": "string",
   "OutputConfig": {
      "KmsKeyId": "string",
      "S3OutputPath": "string",
      "SnsTopicArn": "string"
   },
   "RoleArn": "string",
   "StoppingConditions": {
      "MaxHumanLabeledObjectCount": number,
      "MaxPercentageOfInputDatasetLabeled": number
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HumanTaskConfig

Configures the labeling task and how it is presented to workers; including, but not limited to price, keywords, and batch size (task count).

Type: HumanTaskConfig object

Required: Yes

InputConfig

Input data for the labeling job, such as the Amazon S3 location of the data objects and the location of the manifest file that describes the data objects.

You must specify at least one of the following: `S3DataSource` or `SnsDataSource`.

- Use `SnsDataSource` to specify an SNS input topic for a streaming labeling job. If you do not specify and SNS input topic ARN, Ground Truth will create a one-time labeling job that stops after all data objects in the input manifest file have been labeled.
- Use `S3DataSource` to specify an input manifest file for both streaming and one-time labeling jobs. Adding an `S3DataSource` is optional if you use `SnsDataSource` to create a streaming labeling job.

If you use the Amazon Mechanical Turk workforce, your input data should not include confidential information, personal information or protected health information. Use `ContentClassifiers` to specify that your data is free of personally identifiable information and adult content.

Type: LabelingJobInputConfig object

Required: Yes

LabelAttributeName

The attribute name to use for the label in the output manifest file. This is the key for the key/value pair formed with the label that a worker assigns to the object. The `LabelAttributeName` must meet the following requirements.

- The name can't end with "-metadata".
- If you are using one of the built-in task types or one of the following, the attribute name *must* end with "-ref".
  - Image semantic segmentation (`SemanticSegmentation`) and adjustment (`AdjustmentSemanticSegmentation`) labeling jobs for this task type. One exception is that verification (`VerificationSemanticSegmentation`) *must not* end with -"ref".
  - Video frame object detection (`VideoObjectDetection`), and adjustment and verification (`AdjustmentVideoObjectDetection`) labeling jobs for this task type.
  - Video frame object tracking (`VideoObjectTracking`), and adjustment and verification (`AdjustmentVideoObjectTracking`) labeling jobs for this task type.
  - 3D point cloud semantic segmentation (`3DPointCloudSemanticSegmentation`), and adjustment and verification (`Adjustment3DPointCloudSemanticSegmentation`) labeling jobs for this task type.

- 3D point cloud object tracking (`3DPointCloudObjectTracking`), and adjustment and verification (`Adjustment3DPointCloudObjectTracking`) labeling jobs for this task type.

> **Important**
>
> If you are creating an adjustment or verification labeling job, you must use a *different* `LabelAttributeName` than the one used in the original labeling job. The original labeling job is the Ground Truth labeling job that produced the labels that you want verified or adjusted. To learn more about adjustment and verification labeling jobs, see Verify and Adjust Labels.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 127.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,126}`

Required: Yes

LabelCategoryConfigS3Uri

The S3 URI of the file, referred to as a *label category configuration file*, that defines the categories used to label the data objects.

For 3D point cloud and video frame task types, you can add label category attributes and frame attributes to your label category configuration file. To learn how, see Create a Labeling Category Configuration File for 3D Point Cloud Labeling Jobs.

For named entity recognition jobs, in addition to `"labels"`, you must provide worker instructions in the label category configuration file using the `"instructions"` parameter: `"instructions": {"shortInstruction":"<h1>Add header</h1><p>Add Instructions</p>", "fullInstruction":"<p>Add additional instructions.</p>"}`. For details and an example, see Create a Named Entity Recognition Labeling Job (API) .

For all other built-in task types and custom tasks, your label category configuration file must be a JSON file in the following format. Identify the labels you want to use by replacing `label_1`, `label_2`,`...`,`label_n` with your label categories.

```
{

"document-version": "2018-11-28",

"labels": [{"label": "label_1"},{"label": "label_2"},...{"label":
"label_n"}]

}
```

Note the following about the label category configuration file:

- For image classification and text classification (single and multi-label) you must specify at least two label categories. For all other task types, the minimum number of label categories required is one.
- Each label category must be unique, you cannot specify duplicate label categories.
- If you create a 3D point cloud or video frame adjustment or verification labeling job, you must include `auditLabelAttributeName` in the label category configuration. Use this parameter to enter the `LabelAttributeName` of the labeling job you want to adjust or verify annotations of.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

LabelingJobAlgorithmsConfig

Configures the information required to perform automated data labeling.

Type: LabelingJobAlgorithmsConfig object

Required: No

LabelingJobName

The name of the labeling job. This name is used to identify the job in a list of labeling jobs. Labeling job names must be unique within an AWS account and region. `LabelingJobName` is not case sensitive. For example, Example-job and example-job are considered the same labeling job name by Ground Truth.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

OutputConfig

The location of the output data and the AWS Key Management Service key ID for the key used to encrypt the output data, if any.

Type: LabelingJobOutputConfig object

Required: Yes

RoleArn

The Amazon Resource Number (ARN) that Amazon SageMaker assumes to perform tasks on your behalf during data labeling. You must grant this role the necessary permissions so that Amazon SageMaker can successfully complete data labeling.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

StoppingConditions

A set of conditions for stopping the labeling job. If any of the conditions are met, the job is automatically stopped. You can use these conditions to control the cost of data labeling.

Type: LabelingJobStoppingConditions object

Required: No

Tags

An array of key/value pairs. For more information, see Using Cost Allocation Tags in the *AWS Billing and Cost Management User Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "LabelingJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

LabelingJobArn

The Amazon Resource Name (ARN) of the labeling job. You use this ARN to identify the labeling job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:labeling-job/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateMlflowApp

Service: Amazon SageMaker Service

Creates an MLflow Tracking Server using a general purpose Amazon S3 bucket as the artifact store.

Request Syntax

```
{
   "AccountDefaultStatus": "string",
   "ArtifactStoreUri": "string",
   "DefaultDomainIdList": [ "string" ],
   "ModelRegistrationMode": "string",
   "Name": "string",
   "RoleArn": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "WeeklyMaintenanceWindowStart": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AccountDefaultStatus

Indicates whether this MLflow app is the default for the entire account.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

ArtifactStoreUri

The S3 URI for a general purpose bucket to use as the MLflow App artifact store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

DefaultDomainIdList

List of SageMaker domain IDs for which this MLflow App is used as the default.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: No

ModelRegistrationMode

Whether to enable or disable automatic registration of new MLflow models to the SageMaker Model Registry. To enable automatic model registration, set this value to `AutoModelRegistrationEnabled`. To disable automatic model registration, set this value to `AutoModelRegistrationDisabled`. If not specified, `AutomaticModelRegistration` defaults to `AutoModelRegistrationDisabled`.

Type: String

Valid Values: `AutoModelRegistrationEnabled | AutoModelRegistrationDisabled`

Required: No

Name

A string identifying the MLflow app name. This string is not part of the tracking server ARN.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: Yes

RoleArn

The Amazon Resource Name (ARN) for an IAM role in your account that the MLflow App uses to access the artifact store in Amazon S3. The role should have the `AmazonS3FullAccess` permission.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Tags

Tags consisting of key-value pairs used to manage metadata for the MLflow App.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

WeeklyMaintenanceWindowStart

The day and time of the week in Coordinated Universal Time (UTC) 24-hour standard time that weekly maintenance updates are scheduled. For example: TUE:03:30.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 9.

Pattern: `(Mon|Tue|Wed|Thu|Fri|Sat|Sun):([01]\d|2[0-3]):([0-5]\d)`

Required: No

Response Syntax

```
{
   "Arn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The ARN of the MLflow App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-app/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateMlflowTrackingServer

Service: Amazon SageMaker Service

Creates an MLflow Tracking Server using a general purpose Amazon S3 bucket as the artifact store. For more information, see Create an MLflow Tracking Server.

Request Syntax

```
{
   "ArtifactStoreUri": "string",
   "AutomaticModelRegistration": boolean,
   "MlflowVersion": "string",
   "RoleArn": "string",
   "S3BucketOwnerAccountId": "string",
   "S3BucketOwnerVerification": boolean,
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "TrackingServerName": "string",
   "TrackingServerSize": "string",
   "WeeklyMaintenanceWindowStart": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ArtifactStoreUri

The S3 URI for a general purpose bucket to use as the MLflow Tracking Server artifact store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

AutomaticModelRegistration

Whether to enable or disable automatic registration of new MLflow models to the SageMaker Model Registry. To enable automatic model registration, set this value to `True`. To disable automatic model registration, set this value to `False`. If not specified, `AutomaticModelRegistration` defaults to `False`.

Type: Boolean

Required: No

MlflowVersion

The version of MLflow that the tracking server uses. To see which MLflow versions are available to use, see How it works.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 16.

Pattern: `[0-9]*.[0-9]*.[0-9]*`

Required: No

RoleArn

The Amazon Resource Name (ARN) for an IAM role in your account that the MLflow Tracking Server uses to access the artifact store in Amazon S3. The role should have `AmazonS3FullAccess` permissions. For more information on IAM permissions for tracking server creation, see Set up IAM permissions for MLflow.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

S3BucketOwnerAccountId

Expected AWS account ID that owns the Amazon S3 bucket for artifact storage. Defaults to caller's account ID if not provided.

Type: String

Length Constraints: Fixed length of 12.

Pattern: \d+

Required: No

S3BucketOwnerVerification

Enable Amazon S3 Ownership checks when interacting with Amazon S3 buckets from a SageMaker Managed MLflow Tracking Server. Defaults to `True` if not provided.

Type: Boolean

Required: No

Tags

Tags consisting of key-value pairs used to manage metadata for the tracking server.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TrackingServerName

A unique string identifying the tracking server name. This string is part of the tracking server ARN.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: Yes

TrackingServerSize

The size of the tracking server you want to create. You can choose between `"Small"`, `"Medium"`, and `"Large"`. The default MLflow Tracking Server configuration size is `"Small"`. You can choose a size depending on the projected use of the tracking server such as the volume of data logged, number of users, and frequency of use.

We recommend using a small tracking server for teams of up to 25 users, a medium tracking server for teams of up to 50 users, and a large tracking server for teams of up to 100 users.

Type: String

Valid Values: `Small | Medium | Large`

Required: No

WeeklyMaintenanceWindowStart

The day and time of the week in Coordinated Universal Time (UTC) 24-hour standard time that weekly maintenance updates are scheduled. For example: TUE:03:30.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 9.

Pattern: `(Mon|Tue|Wed|Thu|Fri|Sat|Sun):([01]\d|2[0-3]):([0-5]\d)`

Required: No

Response Syntax

```
{
   "TrackingServerArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrackingServerArn

The ARN of the tracking server.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-tracking-server/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateModel

Service: Amazon SageMaker Service

Creates a model in SageMaker. In the request, you name the model and describe a primary container. For the primary container, you specify the Docker image that contains inference code, artifacts (from prior training), and a custom environment map that the inference code uses when you deploy the model for predictions.

Use this API to create a model if you want to use SageMaker hosting services or run a batch transform job.

To host your model, you create an endpoint configuration with the `CreateEndpointConfig` API, and then create an endpoint with the `CreateEndpoint` API. SageMaker then deploys all of the containers that you defined for the model in the hosting environment.

To run a batch transform using your model, you start a job with the `CreateTransformJob` API. SageMaker uses your model and your dataset to get inferences which are then saved to a specified S3 location.

In the request, you also provide an IAM role that SageMaker can assume to access model artifacts and docker image for deployment on ML compute hosting instances or for batch transform jobs. In addition, you also use the IAM role to manage permissions the inference code needs. For example, if the inference code access any other AWS resources, you grant necessary permissions via this role.

Request Syntax

```
{
   "Containers": [
      {
         "AdditionalModelDataSources": [
            {
               "ChannelName": "string",
               "S3DataSource": {
                  "CompressionType": "string",
                  "ETag": "string",
                  "HubAccessConfig": {
                     "HubContentArn": "string"
                  },
                  "ManifestEtag": "string",
                  "ManifestS3Uri": "string",
                  "ModelAccessConfig": {
```

```
                    "AcceptEula": boolean
                },
                "S3DataType": "string",
                "S3Uri": "string"
            }
        }
    ],
    "ContainerHostname": "string",
    "ContainerMetricsConfig": {
        "MetricsEndpoints": [
            {
                "MetricPublishFrequencyInSeconds": number,
                "MetricsEndpointPath": "string"
            }
        ]
    },
    "Environment": {
        "string" : "string"
    },
    "Image": "string",
    "ImageConfig": {
        "RepositoryAccessMode": "string",
        "RepositoryAuthConfig": {
            "RepositoryCredentialsProviderArn": "string"
        }
    },
    "InferenceSpecificationName": "string",
    "Mode": "string",
    "ModelDataSource": {
        "S3DataSource": {
            "CompressionType": "string",
            "ETag": "string",
            "HubAccessConfig": {
                "HubContentArn": "string"
            },
            "ManifestEtag": "string",
            "ManifestS3Uri": "string",
            "ModelAccessConfig": {
                "AcceptEula": boolean
            },
            "S3DataType": "string",
            "S3Uri": "string"
        }
    },
```

```
         "ModelDataUrl": "string",
         "ModelPackageName": "string",
         "MultiModelConfig": {
            "ModelCacheSetting": "string"
         }
      }
   ],
   "EnableNetworkIsolation": boolean,
   "ExecutionRoleArn": "string",
   "InferenceExecutionConfig": {
      "Mode": "string"
   },
   "ModelName": "string",
   "PrimaryContainer": {
      "AdditionalModelDataSources": [
         {
            "ChannelName": "string",
            "S3DataSource": {
               "CompressionType": "string",
               "ETag": "string",
               "HubAccessConfig": {
                  "HubContentArn": "string"
               },
               "ManifestEtag": "string",
               "ManifestS3Uri": "string",
               "ModelAccessConfig": {
                  "AcceptEula": boolean
               },
               "S3DataType": "string",
               "S3Uri": "string"
            }
         }
      ],
      "ContainerHostname": "string",
      "ContainerMetricsConfig": {
         "MetricsEndpoints": [
            {
               "MetricPublishFrequencyInSeconds": number,
               "MetricsEndpointPath": "string"
            }
         ]
      },
      "Environment": {
         "string" : "string"
```

```
      },
      "Image": "string",
      "ImageConfig": {
         "RepositoryAccessMode": "string",
         "RepositoryAuthConfig": {
            "RepositoryCredentialsProviderArn": "string"
         }
      },
      "InferenceSpecificationName": "string",
      "Mode": "string",
      "ModelDataSource": {
         "S3DataSource": {
            "CompressionType": "string",
            "ETag": "string",
            "HubAccessConfig": {
               "HubContentArn": "string"
            },
            "ManifestEtag": "string",
            "ManifestS3Uri": "string",
            "ModelAccessConfig": {
               "AcceptEula": boolean
            },
            "S3DataType": "string",
            "S3Uri": "string"
         }
      },
      "ModelDataUrl": "string",
      "ModelPackageName": "string",
      "MultiModelConfig": {
         "ModelCacheSetting": "string"
      }
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "VpcConfig": {
      "SecurityGroupIds": [ "string" ],
      "Subnets": [ "string" ]
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Containers

Specifies the containers in the inference pipeline.

Type: Array of ContainerDefinition objects

Array Members: Minimum number of 0 items. Maximum number of 15 items.

Required: No

EnableNetworkIsolation

Isolates the model container. No inbound or outbound network calls can be made to or from the model container.

Type: Boolean

Required: No

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that SageMaker can assume to access model artifacts and docker image for deployment on ML compute instances or for batch transform jobs. Deploying on ML compute instances is part of model hosting. For more information, see SageMaker Roles.

> **Note**
>
> To be able to pass this role to SageMaker, the caller of this API must have the `iam:PassRole` permission.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

InferenceExecutionConfig

Specifies details of how containers in a multi-container endpoint are called.

Type: InferenceExecutionConfig object

Required: No

ModelName

The name of the new model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

PrimaryContainer

The location of the primary docker image containing inference code, associated artifacts, and custom environment map that the inference code uses when the model is deployed for predictions.

Type: ContainerDefinition object

Required: No

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

VpcConfig

A VpcConfig object that specifies the VPC that you want your model to connect to. Control access to and from your model container by configuring the VPC. `VpcConfig` is used in hosting

services and in batch transform. For more information, see Protect Endpoints by Using an Amazon Virtual Private Cloud and Protect Data in Batch Transform Jobs by Using an Amazon Virtual Private Cloud.

Type: VpcConfig object

Required: No

Response Syntax

```
{
   "ModelArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelArn

The ARN of the model created in SageMaker.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:model/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateModelBiasJobDefinition

Service: Amazon SageMaker Service

Creates the definition for a model bias job.

Request Syntax

```
{
   "JobDefinitionName": "string",
   "JobResources": {
      "ClusterConfig": {
         "InstanceCount": number,
         "InstanceType": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
      }
   },
   "ModelBiasAppSpecification": {
      "ConfigUri": "string",
      "Environment": {
         "string" : "string"
      },
      "ImageUri": "string"
   },
   "ModelBiasBaselineConfig": {
      "BaseliningJobName": "string",
      "ConstraintsResource": {
         "S3Uri": "string"
      }
   },
   "ModelBiasJobInput": {
      "BatchTransformInput": {
         "DataCapturedDestinationS3Uri": "string",
         "DatasetFormat": {
            "Csv": {
               "Header": boolean
            },
            "Json": {
               "Line": boolean
            },
            "Parquet": {
            }
         },
```

```
            "EndTimeOffset": "string",
            "ExcludeFeaturesAttribute": "string",
            "FeaturesAttribute": "string",
            "InferenceAttribute": "string",
            "LocalPath": "string",
            "ProbabilityAttribute": "string",
            "ProbabilityThresholdAttribute": number,
            "S3DataDistributionType": "string",
            "S3InputMode": "string",
            "StartTimeOffset": "string"
         },
         "EndpointInput": {
            "EndpointName": "string",
            "EndTimeOffset": "string",
            "ExcludeFeaturesAttribute": "string",
            "FeaturesAttribute": "string",
            "InferenceAttribute": "string",
            "LocalPath": "string",
            "ProbabilityAttribute": "string",
            "ProbabilityThresholdAttribute": number,
            "S3DataDistributionType": "string",
            "S3InputMode": "string",
            "StartTimeOffset": "string"
         },
         "GroundTruthS3Input": {
            "S3Uri": "string"
         }
      },
      "ModelBiasJobOutputConfig": {
         "KmsKeyId": "string",
         "MonitoringOutputs": [
            {
               "S3Output": {
                  "LocalPath": "string",
                  "S3UploadMode": "string",
                  "S3Uri": "string"
               }
            }
         ]
      },
      "NetworkConfig": {
         "EnableInterContainerTrafficEncryption": boolean,
         "EnableNetworkIsolation": boolean,
         "VpcConfig": {
```

```
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
   },
   "RoleArn": "string",
   "StoppingCondition": {
      "MaxRuntimeInSeconds": number
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobDefinitionName

The name of the bias job definition. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

JobResources

Identifies the resources to deploy for a monitoring job.

Type: MonitoringResources object

Required: Yes

ModelBiasAppSpecification

Configures the model bias job to run a specified Docker container image.

Type: ModelBiasAppSpecification object

Required: Yes

ModelBiasBaselineConfig

The baseline configuration for a model bias job.

Type: ModelBiasBaselineConfig object

Required: No

ModelBiasJobInput

Inputs for the model bias job.

Type: ModelBiasJobInput object

Required: Yes

ModelBiasJobOutputConfig

The output configuration for monitoring jobs.

Type: MonitoringOutputConfig object

Required: Yes

NetworkConfig

Networking options for a model bias job.

Type: MonitoringNetworkConfig object

Required: No

RoleArn

The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker AI can assume to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

StoppingCondition

A time limit for how long the monitoring job is allowed to run before stopping.

Type: MonitoringStoppingCondition object

Required: No

Tags

(Optional) An array of key-value pairs. For more information, see Using Cost Allocation Tags in the *AWS Billing and Cost Management User Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "JobDefinitionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobDefinitionArn

The Amazon Resource Name (ARN) of the model bias job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateModelCard

Service: Amazon SageMaker Service

Creates an Amazon SageMaker Model Card.

For information about how to use model cards, see Amazon SageMaker Model Card.

Request Syntax

```
{
   "Content": "string",
   "ModelCardName": "string",
   "ModelCardStatus": "string",
   "SecurityConfig": {
      "KmsKeyId": "string"
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Content

The content of the model card. Content must be in model card JSON schema and provided as a string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 100000.

Pattern: .*

Required: Yes

ModelCardName

The unique name of the model card.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelCardStatus

The approval status of the model card within your organization. Different organizations might have different criteria for model card review and approval.

- `Draft`: The model card is a work in progress.
- `PendingReview`: The model card is pending review.
- `Approved`: The model card is approved.
- `Archived`: The model card is archived. No more updates should be made to the model card, but it can still be exported.

Type: String

Valid Values: `Draft | PendingReview | Approved | Archived`

Required: Yes

SecurityConfig

An optional Key Management Service key to encrypt, decrypt, and re-encrypt model card content for regulated workloads with highly sensitive data.

Type: ModelCardSecurityConfig object

Required: No

Tags

Key-value pairs used to manage metadata for model cards.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "ModelCardArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelCardArn

The Amazon Resource Name (ARN) of the successfully created model card.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-card/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateModelCardExportJob

Service: Amazon SageMaker Service

Creates an Amazon SageMaker Model Card export job.

Request Syntax

```
{
   "ModelCardExportJobName": "string",
   "ModelCardName": "string",
   "ModelCardVersion": number,
   "OutputConfig": {
      "S3OutputPath": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelCardExportJobName

The name of the model card export job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelCardName

The name or Amazon Resource Name (ARN) of the model card to export.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:model-card/.*)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

ModelCardVersion

The version of the model card to export. If a version is not provided, then the latest version of the model card is exported.

Type: Integer

Required: No

OutputConfig

The model card output configuration that specifies the Amazon S3 path for exporting.

Type: ModelCardExportOutputConfig object

Required: Yes

Response Syntax

```
{
   "ModelCardExportJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelCardExportJobArn

The Amazon Resource Name (ARN) of the model card export job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-card/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}/export-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateModelExplainabilityJobDefinition

Service: Amazon SageMaker Service

Creates the definition for a model explainability job.

Request Syntax

```
{
   "JobDefinitionName": "string",
   "JobResources": {
      "ClusterConfig": {
         "InstanceCount": number,
         "InstanceType": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
      }
   },
   "ModelExplainabilityAppSpecification": {
      "ConfigUri": "string",
      "Environment": {
         "string" : "string"
      },
      "ImageUri": "string"
   },
   "ModelExplainabilityBaselineConfig": {
      "BaseliningJobName": "string",
      "ConstraintsResource": {
         "S3Uri": "string"
      }
   },
   "ModelExplainabilityJobInput": {
      "BatchTransformInput": {
         "DataCapturedDestinationS3Uri": "string",
         "DatasetFormat": {
            "Csv": {
               "Header": boolean
            },
            "Json": {
               "Line": boolean
            },
            "Parquet": {
            }
         },
```

```
            "EndTimeOffset": "string",
            "ExcludeFeaturesAttribute": "string",
            "FeaturesAttribute": "string",
            "InferenceAttribute": "string",
            "LocalPath": "string",
            "ProbabilityAttribute": "string",
            "ProbabilityThresholdAttribute": number,
            "S3DataDistributionType": "string",
            "S3InputMode": "string",
            "StartTimeOffset": "string"
         },
         "EndpointInput": {
            "EndpointName": "string",
            "EndTimeOffset": "string",
            "ExcludeFeaturesAttribute": "string",
            "FeaturesAttribute": "string",
            "InferenceAttribute": "string",
            "LocalPath": "string",
            "ProbabilityAttribute": "string",
            "ProbabilityThresholdAttribute": number,
            "S3DataDistributionType": "string",
            "S3InputMode": "string",
            "StartTimeOffset": "string"
         }
      },
      "ModelExplainabilityJobOutputConfig": {
         "KmsKeyId": "string",
         "MonitoringOutputs": [
            {
               "S3Output": {
                  "LocalPath": "string",
                  "S3UploadMode": "string",
                  "S3Uri": "string"
               }
            }
         ]
      },
      "NetworkConfig": {
         "EnableInterContainerTrafficEncryption": boolean,
         "EnableNetworkIsolation": boolean,
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
```

```
   },
   "RoleArn": "string",
   "StoppingCondition": {
      "MaxRuntimeInSeconds": number
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobDefinitionName

The name of the model explainability job definition. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

JobResources

Identifies the resources to deploy for a monitoring job.

Type: MonitoringResources object

Required: Yes

ModelExplainabilityAppSpecification

Configures the model explainability job to run a specified Docker container image.

Type: ModelExplainabilityAppSpecification object

Required: Yes

ModelExplainabilityBaselineConfig

The baseline configuration for a model explainability job.

Type: ModelExplainabilityBaselineConfig object

Required: No

ModelExplainabilityJobInput

Inputs for the model explainability job.

Type: ModelExplainabilityJobInput object

Required: Yes

ModelExplainabilityJobOutputConfig

The output configuration for monitoring jobs.

Type: MonitoringOutputConfig object

Required: Yes

NetworkConfig

Networking options for a model explainability job.

Type: MonitoringNetworkConfig object

Required: No

RoleArn

The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker AI can assume to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

StoppingCondition

A time limit for how long the monitoring job is allowed to run before stopping.

Type: MonitoringStoppingCondition object

Required: No

Tags

(Optional) An array of key-value pairs. For more information, see Using Cost Allocation Tags in the *AWS Billing and Cost Management User Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "JobDefinitionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobDefinitionArn

The Amazon Resource Name (ARN) of the model explainability job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateModelPackage

Service: Amazon SageMaker Service

Creates a model package that you can use to create SageMaker models or list on AWS Marketplace, or a versioned model that is part of a model group. Buyers can subscribe to model packages listed on AWS Marketplace to create models in SageMaker.

To create a model package by specifying a Docker container that contains your inference code and the Amazon S3 location of your model artifacts, provide values for `InferenceSpecification`. To create a model from an algorithm resource that you created or subscribed to in AWS Marketplace, provide a value for `SourceAlgorithmSpecification`.

> **Note**
>
> There are two types of model packages:
>
> - Versioned - a model that is part of a model group in the model registry.
> - Unversioned - a model package that is not part of a model group.

Request Syntax

```
{
   "AdditionalInferenceSpecifications": [
      {
         "Containers": [
            {
               "AdditionalModelDataSources": [
                  {
                     "ChannelName": "string",
                     "S3DataSource": {
                        "CompressionType": "string",
                        "ETag": "string",
                        "HubAccessConfig": {
                           "HubContentArn": "string"
                        },
                        "ManifestEtag": "string",
                        "ManifestS3Uri": "string",
                        "ModelAccessConfig": {
                           "AcceptEula": boolean
                        },
```

```
                "S3DataType": "string",
                "S3Uri": "string"
            }
        }
    ],
    "AdditionalS3DataSource": {
        "CompressionType": "string",
        "ETag": "string",
        "S3DataType": "string",
        "S3Uri": "string"
    },
    "BaseModel": {
        "HubContentName": "string",
        "HubContentVersion": "string",
        "RecipeName": "string"
    },
    "ContainerHostname": "string",
    "Environment": {
        "string" : "string"
    },
    "Framework": "string",
    "FrameworkVersion": "string",
    "Image": "string",
    "ImageDigest": "string",
    "IsCheckpoint": boolean,
    "ModelDataETag": "string",
    "ModelDataSource": {
        "S3DataSource": {
            "CompressionType": "string",
            "ETag": "string",
            "HubAccessConfig": {
                "HubContentArn": "string"
            },
            "ManifestEtag": "string",
            "ManifestS3Uri": "string",
            "ModelAccessConfig": {
                "AcceptEula": boolean
            },
            "S3DataType": "string",
            "S3Uri": "string"
        }
    },
    "ModelDataUrl": "string",
    "ModelInput": {
```

```
                "DataInputConfig": "string"
             },
             "NearestModelName": "string",
             "ProductId": "string"
          }
       ],
       "Description": "string",
       "Name": "string",
       "SupportedContentTypes": [ "string" ],
       "SupportedRealtimeInferenceInstanceTypes": [ "string" ],
       "SupportedResponseMIMETypes": [ "string" ],
       "SupportedTransformInstanceTypes": [ "string" ]
    }
 ],
 "CertifyForMarketplace": boolean,
 "ClientToken": "string",
 "CustomerMetadataProperties": {
    "string" : "string"
 },
 "Domain": "string",
 "DriftCheckBaselines": {
    "Bias": {
       "ConfigFile": {
          "ContentDigest": "string",
          "ContentType": "string",
          "S3Uri": "string"
       },
       "PostTrainingConstraints": {
          "ContentDigest": "string",
          "ContentType": "string",
          "S3Uri": "string"
       },
       "PreTrainingConstraints": {
          "ContentDigest": "string",
          "ContentType": "string",
          "S3Uri": "string"
       }
    },
    "Explainability": {
       "ConfigFile": {
          "ContentDigest": "string",
          "ContentType": "string",
          "S3Uri": "string"
       },
```

```
         "Constraints": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
      },
      "ModelDataQuality": {
         "Constraints": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         },
         "Statistics": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
      },
      "ModelQuality": {
         "Constraints": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         },
         "Statistics": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
      }
   },
   "InferenceSpecification": {
      "Containers": [
         {
            "AdditionalModelDataSources": [
               {
                  "ChannelName": "string",
                  "S3DataSource": {
                     "CompressionType": "string",
                     "ETag": "string",
                     "HubAccessConfig": {
                        "HubContentArn": "string"
                     },
                     "ManifestEtag": "string",
```

```
                "ManifestS3Uri": "string",
                "ModelAccessConfig": {
                   "AcceptEula": boolean
                },
                "S3DataType": "string",
                "S3Uri": "string"
             }
          }
       ],
       "AdditionalS3DataSource": {
          "CompressionType": "string",
          "ETag": "string",
          "S3DataType": "string",
          "S3Uri": "string"
       },
       "BaseModel": {
          "HubContentName": "string",
          "HubContentVersion": "string",
          "RecipeName": "string"
       },
       "ContainerHostname": "string",
       "Environment": {
          "string" : "string"
       },
       "Framework": "string",
       "FrameworkVersion": "string",
       "Image": "string",
       "ImageDigest": "string",
       "IsCheckpoint": boolean,
       "ModelDataETag": "string",
       "ModelDataSource": {
          "S3DataSource": {
             "CompressionType": "string",
             "ETag": "string",
             "HubAccessConfig": {
                "HubContentArn": "string"
             },
             "ManifestEtag": "string",
             "ManifestS3Uri": "string",
             "ModelAccessConfig": {
                "AcceptEula": boolean
             },
             "S3DataType": "string",
             "S3Uri": "string"
```

```
            }
         },
         "ModelDataUrl": "string",
         "ModelInput": {
            "DataInputConfig": "string"
         },
         "NearestModelName": "string",
         "ProductId": "string"
      }
   ],
   "SupportedContentTypes": [ "string" ],
   "SupportedRealtimeInferenceInstanceTypes": [ "string" ],
   "SupportedResponseMIMETypes": [ "string" ],
   "SupportedTransformInstanceTypes": [ "string" ]
},
"ManagedStorageType": "string",
"MetadataProperties": {
   "CommitId": "string",
   "GeneratedBy": "string",
   "ProjectId": "string",
   "Repository": "string"
},
"ModelApprovalStatus": "string",
"ModelCard": {
   "ModelCardContent": "string",
   "ModelCardStatus": "string"
},
"ModelLifeCycle": {
   "Stage": "string",
   "StageDescription": "string",
   "StageStatus": "string"
},
"ModelMetrics": {
   "Bias": {
      "PostTrainingReport": {
         "ContentDigest": "string",
         "ContentType": "string",
         "S3Uri": "string"
      },
      "PreTrainingReport": {
         "ContentDigest": "string",
         "ContentType": "string",
         "S3Uri": "string"
      },
```

```
            "Report": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            }
         },
         "Explainability": {
            "Report": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            }
         },
         "ModelDataQuality": {
            "Constraints": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            },
            "Statistics": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            }
         },
         "ModelQuality": {
            "Constraints": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            },
            "Statistics": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            }
         }
      },
      "ModelPackageDescription": "string",
      "ModelPackageGroupName": "string",
      "ModelPackageName": "string",
      "ModelPackageRegistrationType": "string",
      "SamplePayloadUrl": "string",
      "SecurityConfig": {
```

```
      "KmsKeyId": "string"
   },
   "SkipModelValidation": "string",
   "SourceAlgorithmSpecification": {
      "SourceAlgorithms": [
         {
            "AlgorithmName": "string",
            "ModelDataETag": "string",
            "ModelDataSource": {
               "S3DataSource": {
                  "CompressionType": "string",
                  "ETag": "string",
                  "HubAccessConfig": {
                     "HubContentArn": "string"
                  },
                  "ManifestEtag": "string",
                  "ManifestS3Uri": "string",
                  "ModelAccessConfig": {
                     "AcceptEula": boolean
                  },
                  "S3DataType": "string",
                  "S3Uri": "string"
               }
            },
            "ModelDataUrl": "string"
         }
      ]
   },
   "SourceUri": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "Task": "string",
   "ValidationSpecification": {
      "ValidationProfiles": [
         {
            "ProfileName": "string",
            "TransformJobDefinition": {
               "BatchStrategy": "string",
               "Environment": {
                  "string" : "string"
```

```
            },
            "MaxConcurrentTransforms": number,
            "MaxPayloadInMB": number,
            "TransformInput": {
               "CompressionType": "string",
               "ContentType": "string",
               "DataSource": {
                  "S3DataSource": {
                     "S3DataType": "string",
                     "S3Uri": "string"
                  }
               },
               "SplitType": "string"
            },
            "TransformOutput": {
               "Accept": "string",
               "AssembleWith": "string",
               "KmsKeyId": "string",
               "S3OutputPath": "string"
            },
            "TransformResources": {
               "InstanceCount": number,
               "InstanceType": "string",
               "TransformAmiVersion": "string",
               "VolumeKmsKeyId": "string"
            }
         }
      }
   ],
   "ValidationRole": "string"
}
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AdditionalInferenceSpecifications

An array of additional Inference Specification objects. Each additional Inference Specification specifies artifacts based on this model package that can be used on inference endpoints. Generally used with SageMaker Neo to store the compiled artifacts.

Type: Array of AdditionalInferenceSpecificationDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 15 items.

Required: No

CertifyForMarketplace

Whether to certify the model package for listing on AWS Marketplace.

This parameter is optional for unversioned models, and does not apply to versioned models.

Type: Boolean

Required: No

ClientToken

A unique token that guarantees that the call to this API is idempotent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 36.

Pattern: `[a-zA-Z0-9-]+`

Required: No

CustomerMetadataProperties

The metadata properties associated with the model package versions.

Type: String to string map

Map Entries: Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `([\p{L}\p{Z}\p{N}_.:\/=+\-@]*)${1,128}`

Value Length Constraints: Minimum length of 1. Maximum length of 256.

Value Pattern: `([\p{L}\p{Z}\p{N}_.:\/=+\-@]*)${1,256}`

Required: No

Domain

The machine learning domain of your model package and its components. Common machine learning domains include computer vision and natural language processing.

Type: String

Required: No

DriftCheckBaselines

Represents the drift check baselines that can be used when the model monitor is set using the model package. For more information, see the topic on Drift Detection against Previous Baselines in SageMaker Pipelines in the *Amazon SageMaker Developer Guide*.

Type: DriftCheckBaselines object

Required: No

InferenceSpecification

Specifies details about inference jobs that you can run with models based on this model package, including the following information:

- The Amazon ECR paths of containers that contain the inference code and model artifacts.
- The instance types that the model package supports for transform jobs and real-time endpoints used for inference.
- The input and output content formats that the model package supports for inference.

Type: InferenceSpecification object

Required: No

ManagedStorageType

The storage type of the model package.

Type: String

Valid Values: `Restricted`

Required: No

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Required: No

ModelApprovalStatus

Whether the model is approved for deployment.

This parameter is optional for versioned models, and does not apply to unversioned models.

For versioned models, the value of this parameter must be set to `Approved` to deploy the model.

Type: String

Valid Values: `Approved | Rejected | PendingManualApproval`

Required: No

ModelCard

The model card associated with the model package. Since `ModelPackageModelCard` is tied to a model package, it is a specific usage of a model card and its schema is simplified compared to the schema of `ModelCard`. The `ModelPackageModelCard` schema does not include `model_package_details`, and `model_overview` is composed of the `model_creator` and `model_artifact` properties. For more information about the model package model card schema, see Model package model card schema. For more information about the model card associated with the model package, see View the Details of a Model Version.

Type: ModelPackageModelCard object

Required: No

ModelLifeCycle

A structure describing the current state of the model in its life cycle.

Type: ModelLifeCycle object

Required: No

ModelMetrics

A structure that contains model metrics reports.

Type: ModelMetrics object

Required: No

ModelPackageDescription

A description of the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

ModelPackageGroupName

The name or Amazon Resource Name (ARN) of the model package group that this model version belongs to.

This parameter is required for versioned models, and does not apply to unversioned models.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 170.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*\/)?([a-zA-Z0-9]([a-zA-Z0-9-]){0,62})(?<!-)`

Required: No

ModelPackageName

The name of the model package. The name must have 1 to 63 characters. Valid characters are a-z, A-Z, 0-9, and - (hyphen).

This parameter is required for unversioned models. It is not applicable to versioned models.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelPackageRegistrationType

The package registration type of the model package input.

Type: String

Valid Values: `Logged | Registered`

Required: No

SamplePayloadUrl

The Amazon Simple Storage Service (Amazon S3) path where the sample payload is stored. This path must point to a single gzip compressed tar archive (.tar.gz suffix). This archive can hold multiple files that are all equally used in the load test. Each file in the archive must satisfy the size constraints of the InvokeEndpoint call.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

SecurityConfig

The AWS KMS Key ID (`KMSKeyId`) used for encryption of model package information.

Type: ModelPackageSecurityConfig object

Required: No

SkipModelValidation

Indicates if you want to skip model validation.

Type: String

Valid Values: `All` | None

Required: No

SourceAlgorithmSpecification

Details about the algorithm that was used to create the model package.

Type: SourceAlgorithmSpecification object

Required: No

SourceUri

The URI of the source for the model package. If you want to clone a model package, set it to the model package Amazon Resource Name (ARN). If you want to register a model, set it to the model ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{N}\p{P}]{0,1024}`

Required: No

Tags

A list of key value pairs associated with the model. For more information, see Tagging AWS resources in the *AWS General Reference Guide*.

If you supply `ModelPackageGroupName`, your model package belongs to the model group you specify and uses the tags associated with the model group. In this case, you cannot supply a `tag` argument.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Task

The machine learning task your model package accomplishes. Common machine learning tasks include object detection and image classification. The following tasks are supported

by Inference Recommender: `"IMAGE_CLASSIFICATION" | "OBJECT_DETECTION" | "TEXT_GENERATION" |"IMAGE_SEGMENTATION" | "FILL_MASK" | "CLASSIFICATION" | "REGRESSION" | "OTHER"`.

Specify "OTHER" if none of the tasks listed fit your use case.

Type: String

Required: No

ValidationSpecification

Specifies configurations for one or more transform jobs that SageMaker runs to test the model package.

Type: ModelPackageValidationSpecification object

Required: No

Response Syntax

```
{
   "ModelPackageArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelPackageArn

The Amazon Resource Name (ARN) of the new model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateModelPackageGroup

Service: Amazon SageMaker Service

Creates a model group. A model group contains a group of model versions.

Request Syntax

```
{
   "ManagedConfiguration": {
      "ManagedStorageType": "string"
   },
   "ModelPackageGroupDescription": "string",
   "ModelPackageGroupName": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ManagedConfiguration

The managed configuration of the model package group.

Type: ManagedConfiguration object

Required: No

ModelPackageGroupDescription

A description for the model group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

ModelPackageGroupName

The name of the model group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Tags

A list of key value pairs associated with the model group. For more information, see Tagging AWS resources in the *AWS General Reference Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "ModelPackageGroupArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelPackageGroupArn

The Amazon Resource Name (ARN) of the model group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package-group/[\S]{1,2048}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateModelQualityJobDefinition

Service: Amazon SageMaker Service

Creates a definition for a job that monitors model quality and drift. For information about model monitor, see Amazon SageMaker AI Model Monitor.

Request Syntax

```
{
   "JobDefinitionName": "string",
   "JobResources": {
      "ClusterConfig": {
         "InstanceCount": number,
         "InstanceType": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
      }
   },
   "ModelQualityAppSpecification": {
      "ContainerArguments": [ "string" ],
      "ContainerEntrypoint": [ "string" ],
      "Environment": {
         "string" : "string"
      },
      "ImageUri": "string",
      "PostAnalyticsProcessorSourceUri": "string",
      "ProblemType": "string",
      "RecordPreprocessorSourceUri": "string"
   },
   "ModelQualityBaselineConfig": {
      "BaseliningJobName": "string",
      "ConstraintsResource": {
         "S3Uri": "string"
      }
   },
   "ModelQualityJobInput": {
      "BatchTransformInput": {
         "DataCapturedDestinationS3Uri": "string",
         "DatasetFormat": {
            "Csv": {
               "Header": boolean
            },
            "Json": {
```

```
            "Line": boolean
         },
         "Parquet": {
         }
      },
      "EndTimeOffset": "string",
      "ExcludeFeaturesAttribute": "string",
      "FeaturesAttribute": "string",
      "InferenceAttribute": "string",
      "LocalPath": "string",
      "ProbabilityAttribute": "string",
      "ProbabilityThresholdAttribute": number,
      "S3DataDistributionType": "string",
      "S3InputMode": "string",
      "StartTimeOffset": "string"
   },
   "EndpointInput": {
      "EndpointName": "string",
      "EndTimeOffset": "string",
      "ExcludeFeaturesAttribute": "string",
      "FeaturesAttribute": "string",
      "InferenceAttribute": "string",
      "LocalPath": "string",
      "ProbabilityAttribute": "string",
      "ProbabilityThresholdAttribute": number,
      "S3DataDistributionType": "string",
      "S3InputMode": "string",
      "StartTimeOffset": "string"
   },
   "GroundTruthS3Input": {
      "S3Uri": "string"
   }
},
"ModelQualityJobOutputConfig": {
   "KmsKeyId": "string",
   "MonitoringOutputs": [
      {
         "S3Output": {
            "LocalPath": "string",
            "S3UploadMode": "string",
            "S3Uri": "string"
         }
      }
   ]
```

```
   },
   "NetworkConfig": {
      "EnableInterContainerTrafficEncryption": boolean,
      "EnableNetworkIsolation": boolean,
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
   },
   "RoleArn": "string",
   "StoppingCondition": {
      "MaxRuntimeInSeconds": number
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobDefinitionName

The name of the monitoring job definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

JobResources

Identifies the resources to deploy for a monitoring job.

Type: MonitoringResources object

Required: Yes

ModelQualityAppSpecification

The container that runs the monitoring job.

Type: ModelQualityAppSpecification object

Required: Yes

ModelQualityBaselineConfig

Specifies the constraints and baselines for the monitoring job.

Type: ModelQualityBaselineConfig object

Required: No

ModelQualityJobInput

A list of the inputs that are monitored. Currently endpoints are supported.

Type: ModelQualityJobInput object

Required: Yes

ModelQualityJobOutputConfig

The output configuration for monitoring jobs.

Type: MonitoringOutputConfig object

Required: Yes

NetworkConfig

Specifies the network configuration for the monitoring job.

Type: MonitoringNetworkConfig object

Required: No

RoleArn

The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker AI can assume to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

StoppingCondition

A time limit for how long the monitoring job is allowed to run before stopping.

Type: MonitoringStoppingCondition object

Required: No

Tags

(Optional) An array of key-value pairs. For more information, see Using Cost Allocation Tags in the *AWS Billing and Cost Management User Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "JobDefinitionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobDefinitionArn

The Amazon Resource Name (ARN) of the model quality monitoring job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateMonitoringSchedule

Service: Amazon SageMaker Service

Creates a schedule that regularly starts Amazon SageMaker AI Processing Jobs to monitor the data captured for an Amazon SageMaker AI Endpoint.

Request Syntax

```
{
   "MonitoringScheduleConfig": {
      "MonitoringJobDefinition": {
         "BaselineConfig": {
            "BaseliningJobName": "string",
            "ConstraintsResource": {
               "S3Uri": "string"
            },
            "StatisticsResource": {
               "S3Uri": "string"
            }
         },
         "Environment": {
            "string" : "string"
         },
         "MonitoringAppSpecification": {
            "ContainerArguments": [ "string" ],
            "ContainerEntrypoint": [ "string" ],
            "ImageUri": "string",
            "PostAnalyticsProcessorSourceUri": "string",
            "RecordPreprocessorSourceUri": "string"
         },
         "MonitoringInputs": [
            {
               "BatchTransformInput": {
                  "DataCapturedDestinationS3Uri": "string",
                  "DatasetFormat": {
                     "Csv": {
                        "Header": boolean
                     },
                     "Json": {
                        "Line": boolean
                     },
                     "Parquet": {
                     }
```

```
            },
            "EndTimeOffset": "string",
            "ExcludeFeaturesAttribute": "string",
            "FeaturesAttribute": "string",
            "InferenceAttribute": "string",
            "LocalPath": "string",
            "ProbabilityAttribute": "string",
            "ProbabilityThresholdAttribute": number,
            "S3DataDistributionType": "string",
            "S3InputMode": "string",
            "StartTimeOffset": "string"
         },
         "EndpointInput": {
            "EndpointName": "string",
            "EndTimeOffset": "string",
            "ExcludeFeaturesAttribute": "string",
            "FeaturesAttribute": "string",
            "InferenceAttribute": "string",
            "LocalPath": "string",
            "ProbabilityAttribute": "string",
            "ProbabilityThresholdAttribute": number,
            "S3DataDistributionType": "string",
            "S3InputMode": "string",
            "StartTimeOffset": "string"
         }
      }
   ],
   "MonitoringOutputConfig": {
      "KmsKeyId": "string",
      "MonitoringOutputs": [
         {
            "S3Output": {
               "LocalPath": "string",
               "S3UploadMode": "string",
               "S3Uri": "string"
            }
         }
      ]
   },
   "MonitoringResources": {
      "ClusterConfig": {
         "InstanceCount": number,
         "InstanceType": "string",
         "VolumeKmsKeyId": "string",
```

```
                "VolumeSizeInGB": number
            }
         },
         "NetworkConfig": {
            "EnableInterContainerTrafficEncryption": boolean,
            "EnableNetworkIsolation": boolean,
            "VpcConfig": {
               "SecurityGroupIds": [ "string" ],
               "Subnets": [ "string" ]
            }
         },
         "RoleArn": "string",
         "StoppingCondition": {
            "MaxRuntimeInSeconds": number
         }
      },
      "MonitoringJobDefinitionName": "string",
      "MonitoringType": "string",
      "ScheduleConfig": {
         "DataAnalysisEndTime": "string",
         "DataAnalysisStartTime": "string",
         "ScheduleExpression": "string"
      }
   },
   "MonitoringScheduleName": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MonitoringScheduleConfig

The configuration object that specifies the monitoring schedule and defines the monitoring job.

Type: MonitoringScheduleConfig object

Required: Yes

MonitoringScheduleName

The name of the monitoring schedule. The name must be unique within an AWS Region within an AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Tags

(Optional) An array of key-value pairs. For more information, see Using Cost Allocation Tags in the *AWS Billing and Cost Management User Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "MonitoringScheduleArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

MonitoringScheduleArn

The Amazon Resource Name (ARN) of the monitoring schedule.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateNotebookInstance

Service: Amazon SageMaker Service

Creates an SageMaker AI notebook instance. A notebook instance is a machine learning (ML) compute instance running on a Jupyter notebook.

In a `CreateNotebookInstance` request, specify the type of ML compute instance that you want to run. SageMaker AI launches the instance, installs common libraries that you can use to explore datasets for model training, and attaches an ML storage volume to the notebook instance.

SageMaker AI also provides a set of example notebooks. Each notebook demonstrates how to use SageMaker AI with a specific algorithm or with a machine learning framework.

After receiving the request, SageMaker AI does the following:

1. Creates a network interface in the SageMaker AI VPC.
2. (Option) If you specified `SubnetId`, SageMaker AI creates a network interface in your own VPC, which is inferred from the subnet ID that you provide in the input. When creating this network interface, SageMaker AI attaches the security group that you specified in the request to the network interface that it creates in your VPC.
3. Launches an EC2 instance of the type specified in the request in the SageMaker AI VPC. If you specified `SubnetId` of your VPC, SageMaker AI specifies both network interfaces when launching this instance. This enables inbound traffic from your own VPC to the notebook instance, assuming that the security groups allow it.

After creating the notebook instance, SageMaker AI returns its Amazon Resource Name (ARN). You can't change the name of a notebook instance after you create it.

After SageMaker AI creates the notebook instance, you can connect to the Jupyter server and work in Jupyter notebooks. For example, you can write code to explore a dataset that you can use for model training, train a model, host models by creating SageMaker AI endpoints, and validate hosted models.

For more information, see How It Works.

Request Syntax

```
{
   "AcceleratorTypes": [ "string" ],
```

```
   "AdditionalCodeRepositories": [ "string" ],
   "DefaultCodeRepository": "string",
   "DirectInternetAccess": "string",
   "InstanceMetadataServiceConfiguration": {
      "MinimumInstanceMetadataServiceVersion": "string"
   },
   "InstanceType": "string",
   "IpAddressType": "string",
   "KmsKeyId": "string",
   "LifecycleConfigName": "string",
   "NotebookInstanceName": "string",
   "PlatformIdentifier": "string",
   "RoleArn": "string",
   "RootAccess": "string",
   "SecurityGroupIds": [ "string" ],
   "SubnetId": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "VolumeSizeInGB": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AcceleratorTypes

This parameter is no longer supported. Elastic Inference (EI) is no longer available.

This parameter was used to specify a list of EI instance types to associate with this notebook instance.

Type: Array of strings

Valid Values: `ml.eia1.medium | ml.eia1.large | ml.eia1.xlarge | ml.eia2.medium | ml.eia2.large | ml.eia2.xlarge`

Required: No

AdditionalCodeRepositories

An array of up to three Git repositories to associate with the notebook instance. These can be either the names of Git repositories stored as resources in your account, or the URL of Git repositories in AWS CodeCommit or in any other Git repository. These repositories are cloned at the same level as the default repository of your notebook instance. For more information, see Associating Git Repositories with SageMaker AI Notebook Instances.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 3 items.

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `https://([^/]+)/?(.*)$|^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

DefaultCodeRepository

A Git repository to associate with the notebook instance as its default code repository. This can be either the name of a Git repository stored as a resource in your account, or the URL of a Git repository in AWS CodeCommit or in any other Git repository. When you open a notebook instance, it opens in the directory that contains this repository. For more information, see Associating Git Repositories with SageMaker AI Notebook Instances.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `https://([^/]+)/?(.*)$|^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

DirectInternetAccess

Sets whether SageMaker AI provides internet access to the notebook instance. If you set this to `Disabled` this notebook instance is able to access resources only in your VPC, and is not be able to connect to SageMaker AI training and endpoint services unless you configure a NAT Gateway in your VPC.

For more information, see Notebook Instances Are Internet-Enabled by Default. You can set the value of this parameter to `Disabled` only if you set a value for the `SubnetId` parameter.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

InstanceMetadataServiceConfiguration

Information on the IMDS configuration of the notebook instance

Type: InstanceMetadataServiceConfiguration object

Required: No

InstanceType

The type of ML compute instance to launch for the notebook instance.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.8xlarge | ml.m5d.12xlarge | ml.m5d.16xlarge | ml.m5d.24xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p3dn.24xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge`

| ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p6-b200.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.m6id.large | ml.m6id.xlarge | ml.m6id.2xlarge | ml.m6id.4xlarge | ml.m6id.8xlarge | ml.m6id.12xlarge | ml.m6id.16xlarge | ml.m6id.24xlarge | ml.m6id.32xlarge | ml.c6id.large | ml.c6id.xlarge | ml.c6id.2xlarge | ml.c6id.4xlarge | ml.c6id.8xlarge | ml.c6id.12xlarge | ml.c6id.16xlarge | ml.c6id.24xlarge | ml.c6id.32xlarge | ml.r6id.large | ml.r6id.xlarge | ml.r6id.2xlarge | ml.r6id.4xlarge | ml.r6id.8xlarge | ml.r6id.12xlarge | ml.r6id.16xlarge | ml.r6id.24xlarge | ml.r6id.32xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.p5.4xlarge | ml.p5en.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge

Required: Yes

IpAddressType

The IP address type for the notebook instance. Specify `ipv4` for IPv4-only connectivity or `dualstack` for both IPv4 and IPv6 connectivity. When you specify `dualstack`, the subnet must support IPv6 CIDR blocks. If not specified, defaults to `ipv4`.

Type: String

Valid Values: `ipv4 | dualstack`

Required: No

KmsKeyId

The Amazon Resource Name (ARN) of a AWS Key Management Service key that SageMaker AI uses to encrypt data on the storage volume attached to your notebook instance. The KMS key you provide must be enabled. For information, see Enabling and Disabling Keys in the *AWS Key Management Service Developer Guide*.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

LifecycleConfigName

The name of a lifecycle configuration to associate with the notebook instance. For information about lifestyle configurations, see Step 2.1: (Optional) Customize a Notebook Instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

NotebookInstanceName

The name of the new notebook instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

PlatformIdentifier

The platform identifier of the notebook instance runtime environment. The default value is `notebook-al2023-v1`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 20.

Pattern: `(notebook-al1-v1|notebook-al2-v1|notebook-al2-v2|notebook-al2-v3|notebook-al2023-v1)`

Required: No

RoleArn

When you send any requests to AWS resources from the notebook instance, SageMaker AI assumes this role to perform tasks on your behalf. You must grant this role necessary permissions so SageMaker AI can perform these tasks. The policy must allow the SageMaker AI service principal (sagemaker.amazonaws.com) permissions to assume this role. For more information, see SageMaker AI Roles.

> **Note**
>
> To be able to pass this role to SageMaker AI, the caller of this API must have the `iam:PassRole` permission.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

RootAccess

Whether root access is enabled or disabled for users of the notebook instance. The default value is `Enabled`.

> **Note**
>
> Lifecycle configurations need root access to be able to set up a notebook instance. Because of this, lifecycle configurations associated with a notebook instance always run with root access even if you disable root access for users.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

SecurityGroupIds

The VPC security group IDs, in the form sg-xxxxxxxx. The security groups must be for the same VPC as specified in the subnet.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: No

SubnetId

The ID of the subnet in a VPC to which you would like to have a connectivity from your ML compute instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: No

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

VolumeSizeInGB

The size, in GB, of the ML storage volume to attach to the notebook instance. The default value is 5 GB.

Type: Integer

Valid Range: Minimum value of 5. Maximum value of 16384.

Required: No

Response Syntax

```
{
   "NotebookInstanceArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NotebookInstanceArn

The Amazon Resource Name (ARN) of the notebook instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateNotebookInstanceLifecycleConfig

Service: Amazon SageMaker Service

Creates a lifecycle configuration that you can associate with a notebook instance. A *lifecycle configuration* is a collection of shell scripts that run when you create or start a notebook instance.

Each lifecycle configuration script has a limit of 16384 characters.

The value of the $PATH environment variable that is available to both scripts is `/sbin:bin:/usr/sbin:/usr/bin`.

View Amazon CloudWatch Logs for notebook instance lifecycle configurations in log group `/aws/sagemaker/NotebookInstances` in log stream `[notebook-instance-name]/[LifecycleConfigHook]`.

Lifecycle configuration scripts cannot run for longer than 5 minutes. If a script runs for longer than 5 minutes, it fails and the notebook instance is not created or started.

For information about notebook instance lifestyle configurations, see Step 2.1: (Optional) Customize a Notebook Instance.

> **Note**
>
> Lifecycle configuration scripts execute with root access and the notebook instance's IAM execution role privileges. Grant this permission only to trusted principals. See Customize a Notebook Instance Using a Lifecycle Configuration Script for security best practices.

Request Syntax

```
{
   "NotebookInstanceLifecycleConfigName": "string",
   "OnCreate": [
      {
         "Content": "string"
      }
   ],
   "OnStart": [
      {
         "Content": "string"
```

```
      }
   ],
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

NotebookInstanceLifecycleConfigName

The name of the lifecycle configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

OnCreate

A shell script that runs only once, when you create a notebook instance. The shell script must be a base64-encoded string.

Type: Array of NotebookInstanceLifecycleHook objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

OnStart

A shell script that runs every time you start a notebook instance, including when you create the notebook instance. The shell script must be a base64-encoded string.

Type: Array of NotebookInstanceLifecycleHook objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "NotebookInstanceLifecycleConfigArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NotebookInstanceLifecycleConfigArn

The Amazon Resource Name (ARN) of the lifecycle configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateOptimizationJob

Service: Amazon SageMaker Service

Creates a job that optimizes a model for inference performance. To create the job, you provide the location of a source model, and you provide the settings for the optimization techniques that you want the job to apply. When the job completes successfully, SageMaker uploads the new optimized model to the output destination that you specify.

For more information about how to use this action, and about the supported optimization techniques, see Optimize model inference with Amazon SageMaker.

Request Syntax

```
{
   "DeploymentInstanceType": "string",
   "MaxInstanceCount": number,
   "ModelSource": {
      "S3": {
         "ModelAccessConfig": {
            "AcceptEula": boolean
         },
         "S3Uri": "string"
      },
      "SageMakerModel": {
         "ModelName": "string"
      }
   },
   "OptimizationConfigs": [
      { ... }
   ],
   "OptimizationEnvironment": {
      "string" : "string"
   },
   "OptimizationJobName": "string",
   "OutputConfig": {
      "KmsKeyId": "string",
      "S3OutputLocation": "string",
      "SageMakerModel": {
         "ModelName": "string"
      }
   },
   "RoleArn": "string",
```

```
   "StoppingCondition": {
      "MaxPendingTimeInSeconds": number,
      "MaxRuntimeInSeconds": number,
      "MaxWaitTimeInSeconds": number
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "VpcConfig": {
      "SecurityGroupIds": [ "string" ],
      "Subnets": [ "string" ]
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeploymentInstanceType

The type of instance that hosts the optimized model that you create with the optimization job.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge`

Required: Yes

MaxInstanceCount

The maximum number of instances to use for the optimization job.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

ModelSource

The location of the source model to optimize with an optimization job.

Type: OptimizationJobModelSource object

Required: Yes

OptimizationConfigs

Settings for each of the optimization techniques that the job applies.

Type: Array of OptimizationConfig objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: Yes

OptimizationEnvironment

The environment variables to set in the model container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 25 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `(?!\s*$).+`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

OptimizationJobName

A custom name for the new optimization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

OutputConfig

Details for where to store the optimized model that you create with the optimization job.

Type: OptimizationJobOutputConfig object

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of an IAM role that enables Amazon SageMaker AI to perform tasks on your behalf.

During model optimization, Amazon SageMaker AI needs your permission to:

- Read input data from an S3 bucket
- Write model artifacts to an S3 bucket
- Write logs to Amazon CloudWatch Logs
- Publish metrics to Amazon CloudWatch

You grant permissions for all of these tasks to an IAM role. To pass this role to Amazon SageMaker AI, the caller of this API must have the `iam:PassRole` permission. For more information, see Amazon SageMaker AI Roles.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

StoppingCondition

Specifies a limit to how long a job can run. When the job reaches the time limit, SageMaker ends the job. Use this API to cap costs.

To stop a training job, SageMaker sends the algorithm the `SIGTERM` signal, which delays job termination for 120 seconds. Algorithms can use this 120-second window to save the model artifacts, so the results of training are not lost.

The training algorithms provided by SageMaker automatically save the intermediate results of a model training job when possible. This attempt to save artifacts is only a best effort case as model might not be in a state from which it can be saved. For example, if training has just started, the model might not be ready to save. When saved, this intermediate data is a valid model artifact. You can use it to create a model with `CreateModel`.

> **Note**
>
> The Neural Topic Model (NTM) currently does not support saving intermediate model artifacts. When training NTMs, make sure that the maximum runtime is sufficient for the training job to complete.

Type: StoppingCondition object

Required: Yes

Tags

A list of key-value pairs associated with the optimization job. For more information, see Tagging AWS resources in the *AWS General Reference Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

VpcConfig

A VPC in Amazon VPC that your optimized model has access to.

Type: OptimizationVpcConfig object

Required: No

Response Syntax

```
{
   "OptimizationJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

OptimizationJobArn

The Amazon Resource Name (ARN) of the optimization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:optimization-job/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreatePartnerApp

Service: Amazon SageMaker Service

Creates an Amazon SageMaker Partner AI App.

Request Syntax

```
{
   "ApplicationConfig": {
      "AdminUsers": [ "string" ],
      "Arguments": {
         "string" : "string"
      },
      "AssignedGroupPatterns": [ "string" ],
      "RoleGroupAssignments": [
         {
            "GroupPatterns": [ "string" ],
            "RoleName": "string"
         }
      ]
   },
   "AuthType": "string",
   "ClientToken": "string",
   "EnableAutoMinorVersionUpgrade": boolean,
   "EnableIamSessionBasedIdentity": boolean,
   "ExecutionRoleArn": "string",
   "KmsKeyId": "string",
   "MaintenanceConfig": {
      "MaintenanceWindowStart": "string"
   },
   "Name": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "Tier": "string",
   "Type": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ApplicationConfig

Configuration settings for the SageMaker Partner AI App.

Type: PartnerAppConfig object

Required: No

AuthType

The authorization type that users use to access the SageMaker Partner AI App.

Type: String

Valid Values: `IAM`

Required: Yes

ClientToken

A unique token that guarantees that the call to this API is idempotent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 36.

Pattern: `[a-zA-Z0-9-]+`

Required: No

EnableAutoMinorVersionUpgrade

When set to `TRUE`, the SageMaker Partner AI App is automatically upgraded to the latest minor version during the next scheduled maintenance window, if one is available. Default is `FALSE`.

Type: Boolean

Required: No

EnableIamSessionBasedIdentity

When set to TRUE, the SageMaker Partner AI App sets the AWS IAM session name or the authenticated IAM user as the identity of the SageMaker Partner AI App user.

Type: Boolean

Required: No

ExecutionRoleArn

The ARN of the IAM role that the partner application uses.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

KmsKeyId

SageMaker Partner AI Apps uses AWS KMS to encrypt data at rest using an AWS managed key by default. For more control, specify a customer managed key.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

MaintenanceConfig

Maintenance configuration settings for the SageMaker Partner AI App.

Type: PartnerAppMaintenanceConfig object

Required: No

Name

The name to give the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]+`

Required: Yes

Tags

Each tag consists of a key and an optional value. Tag keys must be unique per resource.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Tier

Indicates the instance type and size of the cluster attached to the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `(?!\s*$).+`

Required: Yes

Type

The type of SageMaker Partner AI App to create. Must be one of the following: `lakera-guard`, `comet`, `deepchecks-llm-evaluation`, or `fiddler`.

Type: String

Valid Values: `lakera-guard | comet | deepchecks-llm-evaluation | fiddler`

Required: Yes

Response Syntax

```
{
   "Arn": "string"
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The ARN of the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:partner-app\/app-[A-Z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreatePartnerAppPresignedUrl

Service: Amazon SageMaker Service

Creates a presigned URL to access an Amazon SageMaker Partner AI App.

Request Syntax

```
{
   "Arn": "string",
   "ExpiresInSeconds": number,
   "SessionExpirationDurationInSeconds": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Arn

The ARN of the SageMaker Partner AI App to create the presigned URL for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:partner-app\/app-[A-Z0-9]{12}`

Required: Yes

ExpiresInSeconds

The time that will pass before the presigned URL expires.

Type: Integer

Valid Range: Minimum value of 5. Maximum value of 300.

Required: No

SessionExpirationDurationInSeconds

Indicates how long the Amazon SageMaker Partner AI App session can be accessed for after logging in.

Type: Integer

Valid Range: Minimum value of 1800. Maximum value of 43200.

Required: No

Response Syntax

```
{
   "Url": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Url

The presigned URL that you can use to access the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreatePipeline

Service: Amazon SageMaker Service

Creates a pipeline using a JSON pipeline definition.

Request Syntax

```
{
   "ClientRequestToken": "string",
   "ParallelismConfiguration": {
      "MaxParallelExecutionSteps": number
   },
   "PipelineDefinition": "string",
   "PipelineDefinitionS3Location": {
      "Bucket": "string",
      "ObjectKey": "string",
      "VersionId": "string"
   },
   "PipelineDescription": "string",
   "PipelineDisplayName": "string",
   "PipelineName": "string",
   "RoleArn": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClientRequestToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the operation. An idempotent operation completes no more than one time.

Type: String

Length Constraints: Minimum length of 32. Maximum length of 128.

Required: Yes

ParallelismConfiguration

This is the configuration that controls the parallelism of the pipeline. If specified, it applies to all runs of this pipeline by default.

Type: ParallelismConfiguration object

Required: No

PipelineDefinition

The JSON pipeline definition of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1048576.

Pattern: `.*(?:[ \r\n\t].*)*`

Required: No

PipelineDefinitionS3Location

The location of the pipeline definition stored in Amazon S3. If specified, SageMaker will retrieve the pipeline definition from this location.

Type: PipelineDefinitionS3Location object

Required: No

PipelineDescription

A description of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

PipelineDisplayName

The display name of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: No

PipelineName

The name of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of the role used by the pipeline to access and create resources.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Tags

A list of tags to apply to the created pipeline.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "PipelineArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

PipelineArn

The Amazon Resource Name (ARN) of the created pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreatePresignedDomainUrl

Service: Amazon SageMaker Service

Creates a URL for a specified UserProfile in a Domain. When accessed in a web browser, the user will be automatically signed in to the domain, and granted access to all of the Apps and files associated with the Domain's Amazon Elastic File System volume. This operation can only be called when the authentication mode equals IAM.

The IAM role or user passed to this API defines the permissions to access the app. Once the presigned URL is created, no additional permission is required to access this URL. IAM authorization policies for this API are also enforced for every HTTP request and WebSocket frame that attempts to connect to the app.

You can restrict access to this API and to the URL that it returns to a list of IP addresses, Amazon VPCs or Amazon VPC Endpoints that you specify. For more information, see Connect to Amazon SageMaker AI Studio Through an Interface VPC Endpoint .

Note

- The URL that you get from a call to `CreatePresignedDomainUrl` has a default timeout of 5 minutes. You can configure this value using `ExpiresInSeconds`. If you try to use the URL after the timeout limit expires, you are directed to the AWS console sign-in page.
- The JupyterLab session default expiration time is 12 hours. You can configure this value using SessionExpirationDurationInSeconds.

Request Syntax

```
{
   "DomainId": "string",
   "ExpiresInSeconds": number,
   "LandingUri": "string",
   "SessionExpirationDurationInSeconds": number,
   "SpaceName": "string",
   "UserProfileName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

ExpiresInSeconds

The number of seconds until the pre-signed URL expires. This value defaults to 300.

Type: Integer

Valid Range: Minimum value of 5. Maximum value of 300.

Required: No

LandingUri

The landing page that the user is directed to when accessing the presigned URL. Using this value, users can access Studio or Studio Classic, even if it is not the default experience for the domain. The supported values are:

- `studio::relative/path`: Directs users to the relative path in Studio.
- `app:JupyterServer:relative/path`: Directs users to the relative path in the Studio Classic application.
- `app:JupyterLab:relative/path`: Directs users to the relative path in the JupyterLab application.
- `app:RStudioServerPro:relative/path`: Directs users to the relative path in the RStudio application.
- `app:CodeEditor:relative/path`: Directs users to the relative path in the Code Editor, based on Code-OSS, Visual Studio Code - Open Source application.

- `app:Canvas:relative/path`: Directs users to the relative path in the Canvas application.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Required: No

SessionExpirationDurationInSeconds

The session expiration duration in seconds. This value defaults to 43200.

Type: Integer

Valid Range: Minimum value of 1800. Maximum value of 43200.

Required: No

SpaceName

The name of the space.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

UserProfileName

The name of the UserProfile to sign-in as.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
```

```
   "AuthorizedUrl": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AuthorizedUrl

The presigned URL.

Type: String

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

CreatePresignedMlflowAppUrl

Service: Amazon SageMaker Service

Returns a presigned URL that you can use to connect to the MLflow UI attached to your MLflow App. For more information, see Launch the MLflow UI using a presigned URL.

Request Syntax

```
{
   "Arn": "string",
   "ExpiresInSeconds": number,
   "SessionExpirationDurationInSeconds": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Arn

The ARN of the MLflow App to connect to your MLflow UI.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-app/.*`

Required: Yes

ExpiresInSeconds

The duration in seconds that your presigned URL is valid. The presigned URL can be used only once.

Type: Integer

Valid Range: Minimum value of 5. Maximum value of 300.

Required: No

SessionExpirationDurationInSeconds

The duration in seconds that your presigned URL is valid. The presigned URL can be used only once.

Type: Integer

Valid Range: Minimum value of 1800. Maximum value of 43200.

Required: No

Response Syntax

```
{
   "AuthorizedUrl": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AuthorizedUrl

A presigned URL with an authorization token.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreatePresignedMlflowTrackingServerUrl

Service: Amazon SageMaker Service

Returns a presigned URL that you can use to connect to the MLflow UI attached to your tracking server. For more information, see Launch the MLflow UI using a presigned URL.

Request Syntax

```
{
   "ExpiresInSeconds": number,
   "SessionExpirationDurationInSeconds": number,
   "TrackingServerName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ExpiresInSeconds

The duration in seconds that your presigned URL is valid. The presigned URL can be used only once.

Type: Integer

Valid Range: Minimum value of 5. Maximum value of 300.

Required: No

SessionExpirationDurationInSeconds

The duration in seconds that your MLflow UI session is valid.

Type: Integer

Valid Range: Minimum value of 1800. Maximum value of 43200.

Required: No

TrackingServerName

The name of the tracking server to connect to your MLflow UI.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: Yes

Response Syntax

```
{
   "AuthorizedUrl": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AuthorizedUrl

A presigned URL with an authorization token.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreatePresignedNotebookInstanceUrl

Service: Amazon SageMaker Service

Returns a URL that you can use to connect to the Jupyter server from a notebook instance. In the SageMaker AI console, when you choose Open next to a notebook instance, SageMaker AI opens a new tab showing the Jupyter server home page from the notebook instance. The console uses this API to get the URL and show the page.

The IAM role or user used to call this API defines the permissions to access the notebook instance. Once the presigned URL is created, no additional permission is required to access this URL. IAM authorization policies for this API are also enforced for every HTTP request and WebSocket frame that attempts to connect to the notebook instance.

You can restrict access to this API and to the URL that it returns to a list of IP addresses that you specify. Use the `NotIpAddress` condition operator and the `aws:SourceIP` condition context key to specify the list of IP addresses that you want to have access to the notebook instance. For more information, see Limit Access to a Notebook Instance by IP Address.

> **Note**
>
> The URL that you get from a call to CreatePresignedNotebookInstanceUrl is valid only for 5 minutes. If you try to use the URL after the 5-minute limit expires, you are directed to the AWS console sign-in page.

Request Syntax

```
{
   "NotebookInstanceName": "string",
   "SessionExpirationDurationInSeconds": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

NotebookInstanceName

The name of the notebook instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

SessionExpirationDurationInSeconds

The duration of the session, in seconds. The default is 12 hours.

Type: Integer

Valid Range: Minimum value of 1800. Maximum value of 43200.

Required: No

Response Syntax

```
{
   "AuthorizedUrl": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AuthorizedUrl

A JSON object that contains the URL string.

Type: String

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateProcessingJob

Service: Amazon SageMaker Service

Creates a processing job.

Request Syntax

```
{
   "AppSpecification": {
      "ContainerArguments": [ "string" ],
      "ContainerEntrypoint": [ "string" ],
      "ImageUri": "string"
   },
   "Environment": {
      "string" : "string"
   },
   "ExperimentConfig": {
      "ExperimentName": "string",
      "RunName": "string",
      "TrialComponentDisplayName": "string",
      "TrialName": "string"
   },
   "NetworkConfig": {
      "EnableInterContainerTrafficEncryption": boolean,
      "EnableNetworkIsolation": boolean,
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
   },
   "ProcessingInputs": [
      {
         "AppManaged": boolean,
         "DatasetDefinition": {
            "AthenaDatasetDefinition": {
               "Catalog": "string",
               "Database": "string",
               "KmsKeyId": "string",
               "OutputCompression": "string",
               "OutputFormat": "string",
               "OutputS3Uri": "string",
               "QueryString": "string",
               "WorkGroup": "string"
```

```
            },
            "DataDistributionType": "string",
            "InputMode": "string",
            "LocalPath": "string",
            "RedshiftDatasetDefinition": {
               "ClusterId": "string",
               "ClusterRoleArn": "string",
               "Database": "string",
               "DbUser": "string",
               "KmsKeyId": "string",
               "OutputCompression": "string",
               "OutputFormat": "string",
               "OutputS3Uri": "string",
               "QueryString": "string"
            }
         },
         "InputName": "string",
         "S3Input": {
            "LocalPath": "string",
            "S3CompressionType": "string",
            "S3DataDistributionType": "string",
            "S3DataType": "string",
            "S3InputMode": "string",
            "S3Uri": "string"
         }
      }
   ],
   "ProcessingJobName": "string",
   "ProcessingOutputConfig": {
      "KmsKeyId": "string",
      "Outputs": [
         {
            "AppManaged": boolean,
            "FeatureStoreOutput": {
               "FeatureGroupName": "string"
            },
            "OutputName": "string",
            "S3Output": {
               "LocalPath": "string",
               "S3UploadMode": "string",
               "S3Uri": "string"
            }
         }
      ]
```

```
   },
   "ProcessingResources": {
      "ClusterConfig": {
         "InstanceCount": number,
         "InstanceType": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
      }
   },
   "RoleArn": "string",
   "StoppingCondition": {
      "MaxRuntimeInSeconds": number
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AppSpecification

Configures the processing job to run a specified Docker container image.

Type: AppSpecification object

Required: Yes

Environment

The environment variables to set in the Docker container. Up to 100 key and values entries in the map are supported.

> **Important**
>
> Do not include any security-sensitive information including account access IDs, secrets, or tokens in any environment fields. As part of the shared responsibility model, you

are responsible for any potential exposure, unauthorized access, or compromise of your sensitive data if caused by security-sensitive information included in the request environment variable or plain text fields.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[\S\s]*`

Required: No

ExperimentConfig

Associates a SageMaker job as a trial component with an experiment and trial. Specified when you call the following APIs:

- CreateProcessingJob
- CreateTrainingJob
- CreateTransformJob

Type: ExperimentConfig object

Required: No

NetworkConfig

Networking options for a processing job, such as whether to allow inbound and outbound network calls to and from processing containers, and the VPC subnets and security groups to use for VPC-enabled processing jobs.

Type: NetworkConfig object

Required: No

ProcessingInputs

An array of inputs configuring the data to download into the processing container.

Type: Array of ProcessingInput objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

ProcessingJobName

The name of the processing job. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ProcessingOutputConfig

Output configuration for the processing job.

Type: ProcessingOutputConfig object

Required: No

ProcessingResources

Identifies the resources, ML compute instances, and ML storage volumes to deploy for a processing job. In distributed training, you specify more than one instance.

Type: ProcessingResources object

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker can assume to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

StoppingCondition

The time limit for how long the processing job is allowed to run.

Type: ProcessingStoppingCondition object

Required: No

Tags

(Optional) An array of key-value pairs. For more information, see Using Cost Allocation Tags in the *AWS Billing and Cost Management User Guide*.

> **Important**
>
> Do not include any security-sensitive information including account access IDs, secrets, or tokens in any tags. As part of the shared responsibility model, you are responsible for any potential exposure, unauthorized access, or compromise of your sensitive data if caused by security-sensitive information included in the request tag variable or plain text fields.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "ProcessingJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ProcessingJobArn

The Amazon Resource Name (ARN) of the processing job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:processing-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateProject

Service: Amazon SageMaker Service

Creates a machine learning (ML) project that can contain one or more templates that set up an ML pipeline from training to deploying an approved model.

Request Syntax

```
{
   "ProjectDescription": "string",
   "ProjectName": "string",
   "ServiceCatalogProvisioningDetails": {
      "PathId": "string",
      "ProductId": "string",
      "ProvisioningArtifactId": "string",
      "ProvisioningParameters": [
         {
            "Key": "string",
            "Value": "string"
         }
      ]
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "TemplateProviders": [
      {
         "CfnTemplateProvider": {
            "Parameters": [
               {
                  "Key": "string",
                  "Value": "string"
               }
            ],
            "RoleARN": "string",
            "TemplateName": "string",
            "TemplateURL": "string"
         }
      }
   ]
```

```
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ProjectDescription

A description for the project.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

ProjectName

The name of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

ServiceCatalogProvisioningDetails

The product ID and provisioning artifact ID to provision a service catalog. The provisioning artifact ID will default to the latest provisioning artifact ID of the product, if you don't provide the provisioning artifact ID. For more information, see What is AWS Service Catalog.

Type: ServiceCatalogProvisioningDetails object

Required: No

Tags

An array of key-value pairs that you want to use to organize and track your AWS resource costs. For more information, see Tagging AWS resources in the *AWS General Reference Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TemplateProviders

An array of template provider configurations for creating infrastructure resources for the project.

Type: Array of CreateTemplateProvider objects

Array Members: Fixed number of 1 item.

Required: No

Response Syntax

```
{
   "ProjectArn": "string",
   "ProjectId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ProjectArn

The Amazon Resource Name (ARN) of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:project/[\S]{1,2048}`

ProjectId

The ID of the new project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateSpace

Service: Amazon SageMaker Service

Creates a private space or a space used for real time collaboration in a domain.

Request Syntax

```
{
   "DomainId": "string",
   "OwnershipSettings": {
      "OwnerUserProfileName": "string"
   },
   "SpaceDisplayName": "string",
   "SpaceName": "string",
   "SpaceSettings": {
      "AppType": "string",
      "CodeEditorAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number
            }
         },
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         }
      },
      "CustomFileSystems": [
         { ... }
      ],
      "JupyterLabAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number
            }
         },
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
```

```
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         }
      },
      "JupyterServerAppSettings": {
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "KernelGatewayAppSettings": {
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
```

```
      },
      "RemoteAccess": "string",
      "SpaceManagedResources": "string",
      "SpaceStorageSettings": {
         "EbsStorageSettings": {
            "EbsVolumeSizeInGb": number
         }
      }
   },
   "SpaceSharingSettings": {
      "SharingType": "string"
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainId

The ID of the associated domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

OwnershipSettings

A collection of ownership settings.

Type: OwnershipSettings object

Required: No

SpaceDisplayName

The name of the space that appears in the SageMaker Studio UI.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `(?!\s*$).+`

Required: No

SpaceName

The name of the space.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

SpaceSettings

A collection of space settings.

Type: SpaceSettings object

Required: No

SpaceSharingSettings

A collection of space sharing settings.

Type: SpaceSharingSettings object

Required: No

Tags

Tags to associated with the space. Each tag consists of a key and an optional value. Tag keys must be unique for each resource. Tags are searchable using the `Search` API.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "SpaceArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

SpaceArn

The space's Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:space/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateStudioLifecycleConfig

Service: Amazon SageMaker Service

Creates a new Amazon SageMaker AI Studio Lifecycle Configuration.

Request Syntax

```
{
   "StudioLifecycleConfigAppType": "string",
   "StudioLifecycleConfigContent": "string",
   "StudioLifecycleConfigName": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

StudioLifecycleConfigAppType

The App type that the Lifecycle Configuration is attached to.

Type: String

Valid Values: `JupyterServer | KernelGateway | CodeEditor | JupyterLab`

Required: Yes

StudioLifecycleConfigContent

The content of your Amazon SageMaker AI Studio Lifecycle Configuration script. This content must be base64 encoded.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 16384.

Pattern: `[\S\s]+`

Required: Yes

StudioLifecycleConfigName

The name of the Amazon SageMaker AI Studio Lifecycle Configuration to create.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Tags

Tags to be associated with the Lifecycle Configuration. Each tag consists of a key and an optional value. Tag keys must be unique per resource. Tags are searchable using the Search API.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "StudioLifecycleConfigArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

StudioLifecycleConfigArn

The ARN of your created Lifecycle Configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:studio-lifecycle-config/.*|None)`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateTrainingJob

Service: Amazon SageMaker Service

Starts a model training job. After training completes, SageMaker saves the resulting model artifacts to an Amazon S3 location that you specify.

If you choose to host your model using SageMaker hosting services, you can use the resulting model artifacts as part of the model. You can also use the artifacts in a machine learning service other than SageMaker, provided that you know how to use them for inference.

In the request body, you provide the following:

- `AlgorithmSpecification` - Identifies the training algorithm to use.
- `HyperParameters` - Specify these algorithm-specific parameters to enable the estimation of model parameters during training. Hyperparameters can be tuned to optimize this learning process. For a list of hyperparameters for each training algorithm provided by SageMaker, see Algorithms.

  > **Important**
  >
  > Do not include any security-sensitive information including account access IDs, secrets, or tokens in any hyperparameter fields. As part of the shared responsibility model, you are responsible for any potential exposure, unauthorized access, or compromise of your sensitive data if caused by security-sensitive information included in the request hyperparameter variable or plain text fields.

- `InputDataConfig` - Describes the input required by the training job and the Amazon S3, EFS, or FSx location where it is stored.
- `OutputDataConfig` - Identifies the Amazon S3 bucket where you want SageMaker to save the results of model training.
- `ResourceConfig` - Identifies the resources, ML compute instances, and ML storage volumes to deploy for model training. In distributed training, you specify more than one instance.
- `EnableManagedSpotTraining` - Optimize the cost of training machine learning models by up to 80% by using Amazon EC2 Spot instances. For more information, see Managed Spot Training.
- `RoleArn` - The Amazon Resource Name (ARN) that SageMaker assumes to perform tasks on your behalf during model training. You must grant this role the necessary permissions so that SageMaker can successfully complete model training.

- `StoppingCondition` - To help cap training costs, use `MaxRuntimeInSeconds` to set a time limit for training. Use `MaxWaitTimeInSeconds` to specify how long a managed spot training job has to complete.
- `Environment` - The environment variables to set in the Docker container.

> **Important**
>
> Do not include any security-sensitive information including account access IDs, secrets, or tokens in any environment fields. As part of the shared responsibility model, you are responsible for any potential exposure, unauthorized access, or compromise of your sensitive data if caused by security-sensitive information included in the request environment variable or plain text fields.

- `RetryStrategy` - The number of times to retry the job when the job fails due to an `InternalServerError`.

For more information about SageMaker, see How It Works.

Request Syntax

```
{
   "AlgorithmSpecification": {
      "AlgorithmName": "string",
      "ContainerArguments": [ "string" ],
      "ContainerEntrypoint": [ "string" ],
      "EnableSageMakerMetricsTimeSeries": boolean,
      "MetricDefinitions": [
         {
            "Name": "string",
            "Regex": "string"
         }
      ],
      "TrainingImage": "string",
      "TrainingImageConfig": {
         "TrainingRepositoryAccessMode": "string",
         "TrainingRepositoryAuthConfig": {
            "TrainingRepositoryCredentialsProviderArn": "string"
         }
      },
      "TrainingInputMode": "string"
```

```
    },
    "CheckpointConfig": {
       "LocalPath": "string",
       "S3Uri": "string"
    },
    "DebugHookConfig": {
       "CollectionConfigurations": [
          {
             "CollectionName": "string",
             "CollectionParameters": {
                "string" : "string"
             }
          }
       ],
       "HookParameters": {
          "string" : "string"
       },
       "LocalPath": "string",
       "S3OutputPath": "string"
    },
    "DebugRuleConfigurations": [
       {
          "InstanceType": "string",
          "LocalPath": "string",
          "RuleConfigurationName": "string",
          "RuleEvaluatorImage": "string",
          "RuleParameters": {
             "string" : "string"
          },
          "S3OutputPath": "string",
          "VolumeSizeInGB": number
       }
    ],
    "EnableInterContainerTrafficEncryption": boolean,
    "EnableManagedSpotTraining": boolean,
    "EnableNetworkIsolation": boolean,
    "Environment": {
       "string" : "string"
    },
    "ExperimentConfig": {
       "ExperimentName": "string",
       "RunName": "string",
       "TrialComponentDisplayName": "string",
       "TrialName": "string"
```

```
    },
    "HyperParameters": {
       "string" : "string"
    },
    "InfraCheckConfig": {
       "EnableInfraCheck": boolean
    },
    "InputDataConfig": [
       {
          "ChannelName": "string",
          "CompressionType": "string",
          "ContentType": "string",
          "DataSource": {
             "DatasetSource": {
                "DatasetArn": "string"
             },
             "FileSystemDataSource": {
                "DirectoryPath": "string",
                "FileSystemAccessMode": "string",
                "FileSystemId": "string",
                "FileSystemType": "string"
             },
             "S3DataSource": {
                "AttributeNames": [ "string" ],
                "HubAccessConfig": {
                   "HubContentArn": "string"
                },
                "InstanceGroupNames": [ "string" ],
                "ModelAccessConfig": {
                   "AcceptEula": boolean
                },
                "S3DataDistributionType": "string",
                "S3DataType": "string",
                "S3Uri": "string"
             }
          },
          "InputMode": "string",
          "RecordWrapperType": "string",
          "ShuffleConfig": {
             "Seed": number
          }
       }
    ],
    "MlflowConfig": {
```

```
      "MlflowExperimentName": "string",
      "MlflowResourceArn": "string",
      "MlflowRunName": "string"
   },
   "ModelPackageConfig": {
      "ModelPackageGroupArn": "string",
      "SourceModelPackageArn": "string"
   },
   "OutputDataConfig": {
      "CompressionType": "string",
      "KmsKeyId": "string",
      "S3OutputPath": "string"
   },
   "ProfilerConfig": {
      "DisableProfiler": boolean,
      "ProfilingIntervalInMilliseconds": number,
      "ProfilingParameters": {
         "string" : "string"
      },
      "S3OutputPath": "string"
   },
   "ProfilerRuleConfigurations": [
      {
         "InstanceType": "string",
         "LocalPath": "string",
         "RuleConfigurationName": "string",
         "RuleEvaluatorImage": "string",
         "RuleParameters": {
            "string" : "string"
         },
         "S3OutputPath": "string",
         "VolumeSizeInGB": number
      }
   ],
   "RemoteDebugConfig": {
      "EnableRemoteDebug": boolean
   },
   "ResourceConfig": {
      "InstanceCount": number,
      "InstanceGroups": [
         {
            "InstanceCount": number,
            "InstanceGroupName": "string",
            "InstanceType": "string"
```

```
            }
         ],
         "InstancePlacementConfig": {
            "EnableMultipleJobs": boolean,
            "PlacementSpecifications": [
               {
                  "InstanceCount": number,
                  "UltraServerId": "string"
               }
            ]
         },
         "InstanceType": "string",
         "KeepAlivePeriodInSeconds": number,
         "TrainingPlanArn": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
      },
      "RetryStrategy": {
         "MaximumRetryAttempts": number
      },
      "RoleArn": "string",
      "ServerlessJobConfig": {
         "AcceptEula": boolean,
         "BaseModelArn": "string",
         "CustomizationTechnique": "string",
         "EvaluationType": "string",
         "EvaluatorArn": "string",
         "JobType": "string",
         "Peft": "string"
      },
      "SessionChainingConfig": {
         "EnableSessionTagChaining": boolean
      },
      "StoppingCondition": {
         "MaxPendingTimeInSeconds": number,
         "MaxRuntimeInSeconds": number,
         "MaxWaitTimeInSeconds": number
      },
      "Tags": [
         {
            "Key": "string",
            "Value": "string"
         }
      ],
```

```
   "TensorBoardOutputConfig": {
      "LocalPath": "string",
      "S3OutputPath": "string"
   },
   "TrainingJobName": "string",
   "VpcConfig": {
      "SecurityGroupIds": [ "string" ],
      "Subnets": [ "string" ]
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AlgorithmSpecification

The registry path of the Docker image that contains the training algorithm and algorithm-specific metadata, including the input mode. For more information about algorithms provided by SageMaker, see Algorithms. For information about providing your own algorithms, see Using Your Own Algorithms with Amazon SageMaker.

Type: AlgorithmSpecification object

Required: No

CheckpointConfig

Contains information about the output location for managed spot training checkpoint data.

Type: CheckpointConfig object

Required: No

DebugHookConfig

Configuration information for the Amazon SageMaker Debugger hook parameters, metric and tensor collections, and storage paths. To learn more about how to configure the `DebugHookConfig` parameter, see Use the SageMaker and Debugger Configuration API Operations to Create, Update, and Debug Your Training Job.

Type: DebugHookConfig object

Required: No

DebugRuleConfigurations

Configuration information for Amazon SageMaker Debugger rules for debugging output tensors.

Type: Array of DebugRuleConfiguration objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

EnableInterContainerTrafficEncryption

To encrypt all communications between ML compute instances in distributed training, choose `True`. Encryption provides greater security for distributed training, but training might take longer. How long it takes depends on the amount of communication between compute instances, especially if you use a deep learning algorithm in distributed training. For more information, see Protect Communications Between ML Compute Instances in a Distributed Training Job.

Type: Boolean

Required: No

EnableManagedSpotTraining

To train models using managed spot training, choose `True`. Managed spot training provides a fully managed and scalable infrastructure for training machine learning models. this option is useful when training jobs can be interrupted and when there is flexibility when the training job is run.

The complete and intermediate results of jobs are stored in an Amazon S3 bucket, and can be used as a starting point to train models incrementally. Amazon SageMaker provides metrics and logs in CloudWatch. They can be used to see when managed spot training jobs are running, interrupted, resumed, or completed.

Type: Boolean

Required: No

EnableNetworkIsolation

Isolates the training container. No inbound or outbound network calls can be made, except for calls between peers within a training cluster for distributed training. If you enable network isolation for training jobs that are configured to use a VPC, SageMaker downloads and uploads customer data and model artifacts through the specified VPC, but the training container does not have network access.

Type: Boolean

Required: No

Environment

The environment variables to set in the Docker container.

> **Important**
>
> Do not include any security-sensitive information including account access IDs, secrets, or tokens in any environment fields. As part of the shared responsibility model, you are responsible for any potential exposure, unauthorized access, or compromise of your sensitive data if caused by security-sensitive information included in the request environment variable or plain text fields.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 512.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 512.

Value Pattern: `[\S\s]*`

Required: No

ExperimentConfig

Associates a SageMaker job as a trial component with an experiment and trial. Specified when you call the following APIs:

- CreateProcessingJob
- CreateTrainingJob
- CreateTransformJob

Type: ExperimentConfig object

Required: No

HyperParameters

Algorithm-specific parameters that influence the quality of the model. You set hyperparameters before you start the learning process. For a list of hyperparameters for each training algorithm provided by SageMaker, see Algorithms.

You can specify a maximum of 100 hyperparameters. Each hyperparameter is a key-value pair. Each key and value is limited to 256 characters, as specified by the `Length Constraint`.

> **Important**
>
> Do not include any security-sensitive information including account access IDs, secrets, or tokens in any hyperparameter fields. As part of the shared responsibility model, you are responsible for any potential exposure, unauthorized access, or compromise of your sensitive data if caused by any security-sensitive information included in the request hyperparameter variable or plain text fields.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Required: No

InfraCheckConfig

Contains information about the infrastructure health check configuration for the training job.

Type: InfraCheckConfig object

Required: No

InputDataConfig

An array of `Channel` objects. Each channel is a named input source. `InputDataConfig` describes the input data and its location.

Algorithms can accept input data from one or more channels. For example, an algorithm might have two channels of input data, `training_data` and `validation_data`. The configuration for each channel provides the S3, EFS, or FSx location where the input data is stored. It also provides information about the stored data: the MIME type, compression method, and whether the data is wrapped in RecordIO format.

Depending on the input mode that the algorithm supports, SageMaker either copies input data files from an S3 bucket to a local directory in the Docker container, or makes it available as input streams. For example, if you specify an EFS location, input data files are available as input streams. They do not need to be downloaded.

Your input must be in the same AWS region as your training job.

Type: Array of Channel objects

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Required: No

MlflowConfig

The MLflow configuration using SageMaker managed MLflow.

Type: MlflowConfig object

Required: No

ModelPackageConfig

The configuration for the model package.

Type: ModelPackageConfig object

Required: No

OutputDataConfig

Specifies the path to the S3 location where you want to store model artifacts. SageMaker creates subfolders for the artifacts.

Type: OutputDataConfig object

Required: Yes

ProfilerConfig

Configuration information for Amazon SageMaker Debugger system monitoring, framework profiling, and storage paths.

Type: ProfilerConfig object

Required: No

ProfilerRuleConfigurations

Configuration information for Amazon SageMaker Debugger rules for profiling system and framework metrics.

Type: Array of ProfilerRuleConfiguration objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

RemoteDebugConfig

Configuration for remote debugging. To learn more about the remote debugging functionality of SageMaker, see Access a training container through AWS Systems Manager (SSM) for remote debugging.

Type: RemoteDebugConfig object

Required: No

ResourceConfig

The resources, including the ML compute instances and ML storage volumes, to use for model training.

ML storage volumes store model artifacts and incremental states. Training algorithms might also use ML storage volumes for scratch space. If you want SageMaker to use the ML storage

volume to store the training data, choose `File` as the `TrainingInputMode` in the algorithm specification. For distributed training algorithms, specify an instance count greater than 1.

Type: ResourceConfig object

Required: No

RetryStrategy

The number of times to retry the job when the job fails due to an `InternalServerError`.

Type: RetryStrategy object

Required: No

RoleArn

The Amazon Resource Name (ARN) of an IAM role that SageMaker can assume to perform tasks on your behalf.

During model training, SageMaker needs your permission to read input data from an S3 bucket, download a Docker image that contains training code, write model artifacts to an S3 bucket, write logs to Amazon CloudWatch Logs, and publish metrics to Amazon CloudWatch. You grant permissions for all of these tasks to an IAM role. For more information, see SageMaker Roles.

> **Note**
>
> To be able to pass this role to SageMaker, the caller of this API must have the `iam:PassRole` permission.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

ServerlessJobConfig

The configuration for serverless training jobs.

Type: ServerlessJobConfig object

Required: No

SessionChainingConfig

Contains information about attribute-based access control (ABAC) for the training job.

Type: SessionChainingConfig object

Required: No

StoppingCondition

Specifies a limit to how long a model training job can run. It also specifies how long a managed Spot training job has to complete. When the job reaches the time limit, SageMaker ends the training job. Use this API to cap model training costs.

To stop a job, SageMaker sends the algorithm the `SIGTERM` signal, which delays job termination for 120 seconds. Algorithms can use this 120-second window to save the model artifacts, so the results of training are not lost.

Type: StoppingCondition object

Required: No

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

> **Important**
>
> Do not include any security-sensitive information including account access IDs, secrets, or tokens in any tags. As part of the shared responsibility model, you are responsible for any potential exposure, unauthorized access, or compromise of your sensitive data if caused by any security-sensitive information included in the request tag variable or plain text fields.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TensorBoardOutputConfig

Configuration of storage locations for the Amazon SageMaker Debugger TensorBoard output data.

Type: TensorBoardOutputConfig object

Required: No

TrainingJobName

The name of the training job. The name must be unique within an AWS Region in an AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

VpcConfig

A VpcConfig object that specifies the VPC that you want your training job to connect to. Control access to and from your training container by configuring the VPC. For more information, see Protect Training Jobs by Using an Amazon Virtual Private Cloud.

Type: VpcConfig object

Required: No

Response Syntax

```
{
   "TrainingJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrainingJobArn

The Amazon Resource Name (ARN) of the training job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateTrainingPlan

Service: Amazon SageMaker Service

Creates a new training plan in SageMaker to reserve compute capacity.

Amazon SageMaker Training Plan is a capability within SageMaker that allows customers to reserve and manage GPU capacity for large-scale AI model training. It provides a way to secure predictable access to computational resources within specific timelines and budgets, without the need to manage underlying infrastructure.

How it works

Plans can be created for specific resources such as SageMaker Training Jobs or SageMaker HyperPod clusters, automatically provisioning resources, setting up infrastructure, executing workloads, and handling infrastructure failures.

Plan creation workflow

- Users search for available plan offerings based on their requirements (e.g., instance type, count, start time, duration) using the `SearchTrainingPlanOfferings` API operation.
- They create a plan that best matches their needs using the ID of the plan offering they want to use.
- After successful upfront payment, the plan's status becomes `Scheduled`.
- The plan can be used to:
  - Queue training jobs.
  - Allocate to an instance group of a SageMaker HyperPod cluster.
- When the plan start date arrives, it becomes `Active`. Based on available reserved capacity:
  - Training jobs are launched.
  - Instance groups are provisioned.

Plan composition

A plan can consist of one or more Reserved Capacities, each defined by a specific instance type, quantity, Availability Zone, duration, and start and end times. For more information about Reserved Capacity, see `ReservedCapacitySummary` .

Request Syntax

```
{
   "SpareInstanceCountPerUltraServer": number,
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "TrainingPlanName": "string",
   "TrainingPlanOfferingId": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

SpareInstanceCountPerUltraServer

Number of spare instances to reserve per UltraServer for enhanced resiliency. Default is 1.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

Tags

An array of key-value pairs to apply to this training plan.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TrainingPlanName

The name of the training plan to create.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: Yes

TrainingPlanOfferingId

The unique identifier of the training plan offering to use for creating this plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-z0-9\-]+`

Required: Yes

Response Syntax

```
{
   "TrainingPlanArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrainingPlanArn

The Amazon Resource Name (ARN); of the created training plan.

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateTransformJob

Service: Amazon SageMaker Service

Starts a transform job. A transform job uses a trained model to get inferences on a dataset and saves these results to an Amazon S3 location that you specify.

To perform batch transformations, you create a transform job and use the data that you have readily available.

In the request body, you provide the following:

- `TransformJobName` - Identifies the transform job. The name must be unique within an AWS Region in an AWS account.
- `ModelName` - Identifies the model to use. `ModelName` must be the name of an existing Amazon SageMaker model in the same AWS Region and AWS account. For information on creating a model, see CreateModel.
- `TransformInput` - Describes the dataset to be transformed and the Amazon S3 location where it is stored.
- `TransformOutput` - Identifies the Amazon S3 location where you want Amazon SageMaker to save the results from the transform job.
- `TransformResources` - Identifies the ML compute instances and AMI image versions for the transform job.

For more information about how batch transformation works, see Batch Transform.

Request Syntax

```
{
   "BatchStrategy": "string",
   "DataCaptureConfig": {
      "DestinationS3Uri": "string",
      "GenerateInferenceId": boolean,
      "KmsKeyId": "string"
   },
   "DataProcessing": {
      "InputFilter": "string",
      "JoinSource": "string",
      "OutputFilter": "string"
   },
```

```
   "Environment": {
      "string" : "string"
   },
   "ExperimentConfig": {
      "ExperimentName": "string",
      "RunName": "string",
      "TrialComponentDisplayName": "string",
      "TrialName": "string"
   },
   "MaxConcurrentTransforms": number,
   "MaxPayloadInMB": number,
   "ModelClientConfig": {
      "InvocationsMaxRetries": number,
      "InvocationsTimeoutInSeconds": number
   },
   "ModelName": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "TransformInput": {
      "CompressionType": "string",
      "ContentType": "string",
      "DataSource": {
         "S3DataSource": {
            "S3DataType": "string",
            "S3Uri": "string"
         }
      },
      "SplitType": "string"
   },
   "TransformJobName": "string",
   "TransformOutput": {
      "Accept": "string",
      "AssembleWith": "string",
      "KmsKeyId": "string",
      "S3OutputPath": "string"
   },
   "TransformResources": {
      "InstanceCount": number,
      "InstanceType": "string",
      "TransformAmiVersion": "string",
```

```
        "VolumeKmsKeyId": "string"
    }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

BatchStrategy

Specifies the number of records to include in a mini-batch for an HTTP inference request. A *record* is a single unit of input data that inference can be made on. For example, a single line in a CSV file is a record.

To enable the batch strategy, you must set the `SplitType` property to `Line`, `RecordIO`, or `TFRecord`.

To use only one record when making an HTTP invocation request to a container, set `BatchStrategy` to `SingleRecord` and `SplitType` to `Line`.

To fit as many records in a mini-batch as can fit within the `MaxPayloadInMB` limit, set `BatchStrategy` to `MultiRecord` and `SplitType` to `Line`.

Type: String

Valid Values: `MultiRecord | SingleRecord`

Required: No

DataCaptureConfig

Configuration to control how SageMaker captures inference data.

Type: BatchDataCaptureConfig object

Required: No

DataProcessing

The data structure used to specify the data to be used for inference in a batch transform job and to associate the data that is relevant to the prediction results in the output. The input filter

provided allows you to exclude input data that is not needed for inference in a batch transform job. The output filter provided allows you to include input data relevant to interpreting the predictions in the output from the job. For more information, see Associate Prediction Results with their Corresponding Input Records.

Type: DataProcessing object

Required: No

Environment

The environment variables to set in the Docker container. Don't include any sensitive data in your environment variables. We support up to 16 key and values entries in the map.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 16 items.

Key Length Constraints: Minimum length of 0. Maximum length of 1024.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]{0,1023}`

Value Length Constraints: Minimum length of 0. Maximum length of 10240.

Value Pattern: `[\S\s]*`

Required: No

ExperimentConfig

Associates a SageMaker job as a trial component with an experiment and trial. Specified when you call the following APIs:

- CreateProcessingJob
- CreateTrainingJob
- CreateTransformJob

Type: ExperimentConfig object

Required: No

MaxConcurrentTransforms

The maximum number of parallel requests that can be sent to each instance in a transform job. If `MaxConcurrentTransforms` is set to `0` or left unset, Amazon SageMaker checks the

optional execution-parameters to determine the settings for your chosen algorithm. If the execution-parameters endpoint is not enabled, the default value is 1. For more information on execution-parameters, see How Containers Serve Requests. For built-in algorithms, you don't need to set a value for `MaxConcurrentTransforms`.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

MaxPayloadInMB

The maximum allowed size of the payload, in MB. A *payload* is the data portion of a record (without metadata). The value in `MaxPayloadInMB` must be greater than, or equal to, the size of a single record. To estimate the size of a record in MB, divide the size of your dataset by the number of records. To ensure that the records fit within the maximum payload size, we recommend using a slightly larger value. The default value is 6 MB.

The value of `MaxPayloadInMB` cannot be greater than 100 MB. If you specify the `MaxConcurrentTransforms` parameter, the value of `(MaxConcurrentTransforms * MaxPayloadInMB)` also cannot exceed 100 MB.

For cases where the payload might be arbitrarily large and is transmitted using HTTP chunked encoding, set the value to `0`. This feature works only in supported algorithms. Currently, Amazon SageMaker built-in algorithms do not support HTTP chunked encoding.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

ModelClientConfig

Configures the timeout and maximum number of retries for processing a transform job invocation.

Type: ModelClientConfig object

Required: No

ModelName

The name of the model that you want to use for the transform job. `ModelName` must be the name of an existing Amazon SageMaker model within an AWS Region in an AWS account.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

Tags

(Optional) An array of key-value pairs. For more information, see Using Cost Allocation Tags in the *AWS Billing and Cost Management User Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TransformInput

Describes the input source and the way the transform job consumes it.

Type: TransformInput object

Required: Yes

TransformJobName

The name of the transform job. The name must be unique within an AWS Region in an AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

TransformOutput

Describes the results of the transform job.

Type: TransformOutput object

Required: Yes

TransformResources

Describes the resources, including ML instance types and ML instance count, to use for the transform job.

Type: TransformResources object

Required: Yes

Response Syntax

```
{
   "TransformJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TransformJobArn

The Amazon Resource Name (ARN) of the transform job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:transform-job/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateTrial

Service: Amazon SageMaker Service

Creates an SageMaker *trial*. A trial is a set of steps called *trial components* that produce a machine learning model. A trial is part of a single SageMaker *experiment*.

When you use SageMaker Studio or the SageMaker Python SDK, all experiments, trials, and trial components are automatically tracked, logged, and indexed. When you use the AWS SDK for Python (Boto), you must use the logging APIs provided by the SDK.

You can add tags to a trial and then use the Search API to search for the tags.

To get a list of all your trials, call the ListTrials API. To view a trial's properties, call the DescribeTrial API. To create a trial component, call the CreateTrialComponent API.

Request Syntax

```
{
   "DisplayName": "string",
   "ExperimentName": "string",
   "MetadataProperties": {
      "CommitId": "string",
      "GeneratedBy": "string",
      "ProjectId": "string",
      "Repository": "string"
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "TrialName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DisplayName

The name of the trial as displayed. The name doesn't need to be unique. If `DisplayName` isn't specified, `TrialName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

ExperimentName

The name of the experiment to associate the trial with.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Required: No

Tags

A list of tags to associate with the trial. You can use Search API to search on the tags.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TrialName

The name of the trial. The name must be unique in your AWS account and is not case-sensitive.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "TrialArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrialArn

The Amazon Resource Name (ARN) of the trial.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateTrialComponent

Service: Amazon SageMaker Service

Creates a *trial component*, which is a stage of a machine learning *trial*. A trial is composed of one or more trial components. A trial component can be used in multiple trials.

Trial components include pre-processing jobs, training jobs, and batch transform jobs.

When you use SageMaker Studio or the SageMaker Python SDK, all experiments, trials, and trial components are automatically tracked, logged, and indexed. When you use the AWS SDK for Python (Boto), you must use the logging APIs provided by the SDK.

You can add tags to a trial component and then use the Search API to search for the tags.

Request Syntax

```
{
   "DisplayName": "string",
   "EndTime": number,
   "InputArtifacts": {
      "string" : {
         "MediaType": "string",
         "Value": "string"
      }
   },
   "MetadataProperties": {
      "CommitId": "string",
      "GeneratedBy": "string",
      "ProjectId": "string",
      "Repository": "string"
   },
   "OutputArtifacts": {
      "string" : {
         "MediaType": "string",
         "Value": "string"
      }
   },
   "Parameters": {
      "string" : {
         "NumberValue": number,
         "StringValue": "string"
      }
   },
```

```
   "StartTime": number,
   "Status": {
      "Message": "string",
      "PrimaryStatus": "string"
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "TrialComponentName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DisplayName

The name of the component as displayed. The name doesn't need to be unique. If `DisplayName` isn't specified, `TrialComponentName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

EndTime

When the component ended.

Type: Timestamp

Required: No

InputArtifacts

The input artifacts for the component. Examples of input artifacts are datasets, algorithms, hyperparameters, source code, and instance types.

Type: String to TrialComponentArtifact object map

Map Entries: Minimum number of 0 items. Maximum number of 60 items.

Key Length Constraints: Minimum length of 0. Maximum length of 128.

Key Pattern: .*

Required: No

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Required: No

OutputArtifacts

The output artifacts for the component. Examples of output artifacts are metrics, snapshots, logs, and images.

Type: String to TrialComponentArtifact object map

Map Entries: Minimum number of 0 items. Maximum number of 60 items.

Key Length Constraints: Minimum length of 0. Maximum length of 128.

Key Pattern: .*

Required: No

Parameters

The hyperparameters for the component.

Type: String to TrialComponentParameterValue object map

Map Entries: Minimum number of 0 items. Maximum number of 300 items.

Key Length Constraints: Minimum length of 0. Maximum length of 320.

Key Pattern: .*

Required: No

StartTime

When the component started.

Type: Timestamp

Required: No

Status

The status of the component. States include:

- InProgress
- Completed
- Failed

Type: TrialComponentStatus object

Required: No

Tags

A list of tags to associate with the component. You can use Search API to search on the tags.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TrialComponentName

The name of the component. The name must be unique in your AWS account and is not case-sensitive.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "TrialComponentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrialComponentArn

The Amazon Resource Name (ARN) of the trial component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial-component/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateUserProfile

Service: Amazon SageMaker Service

Creates a user profile. A user profile represents a single user within a domain, and is the main way to reference a "person" for the purposes of sharing, reporting, and other user-oriented features. This entity is created when a user onboards to a domain. If an administrator invites a person by email or imports them from IAM Identity Center, a user profile is automatically created. A user profile is the primary holder of settings for an individual user and has a reference to the user's private Amazon Elastic File System home directory.

Request Syntax

```
{
   "DomainId": "string",
   "SingleSignOnUserIdentifier": "string",
   "SingleSignOnUserValue": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "UserProfileName": "string",
   "UserSettings": {
      "AutoMountHomeEFS": "string",
      "CanvasAppSettings": {
         "DirectDeploySettings": {
            "Status": "string"
         },
         "EmrServerlessSettings": {
            "ExecutionRoleArn": "string",
            "Status": "string"
         },
         "GenerativeAiSettings": {
            "AmazonBedrockRoleArn": "string"
         },
         "IdentityProviderOAuthSettings": [
            {
               "DataSourceName": "string",
               "SecretArn": "string",
               "Status": "string"
            }
```

```
        ],
        "KendraSettings": {
           "Status": "string"
        },
        "ModelRegisterSettings": {
           "CrossAccountModelRegisterRoleArn": "string",
           "Status": "string"
        },
        "TimeSeriesForecastingSettings": {
           "AmazonForecastRoleArn": "string",
           "Status": "string"
        },
        "WorkspaceSettings": {
           "S3ArtifactPath": "string",
           "S3KmsKeyId": "string"
        }
     },
     "CodeEditorAppSettings": {
        "AppLifecycleManagement": {
           "IdleSettings": {
              "IdleTimeoutInMinutes": number,
              "LifecycleManagement": "string",
              "MaxIdleTimeoutInMinutes": number,
              "MinIdleTimeoutInMinutes": number
           }
        },
        "BuiltInLifecycleConfigArn": "string",
        "CustomImages": [
           {
              "AppImageConfigName": "string",
              "ImageName": "string",
              "ImageVersionNumber": number
           }
        ],
        "DefaultResourceSpec": {
           "InstanceType": "string",
           "LifecycleConfigArn": "string",
           "SageMakerImageArn": "string",
           "SageMakerImageVersionAlias": "string",
           "SageMakerImageVersionArn": "string",
           "TrainingPlanArn": "string"
        },
        "LifecycleConfigArns": [ "string" ]
     },
```

```
      "CustomFileSystemConfigs": [
         { ... }
      ],
      "CustomPosixUserConfig": {
         "Gid": number,
         "Uid": number
      },
      "DefaultLandingUri": "string",
      "ExecutionRole": "string",
      "JupyterLabAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number,
               "LifecycleManagement": "string",
               "MaxIdleTimeoutInMinutes": number,
               "MinIdleTimeoutInMinutes": number
            }
         },
         "BuiltInLifecycleConfigArn": "string",
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "EmrSettings": {
            "AssumableRoleArns": [ "string" ],
            "ExecutionRoleArns": [ "string" ]
         },
         "LifecycleConfigArns": [ "string" ]
```

```
      },
      "JupyterServerAppSettings": {
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "KernelGatewayAppSettings": {
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "RSessionAppSettings": {
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
```

```
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         }
      },
      "RStudioServerProAppSettings": {
         "AccessStatus": "string",
         "UserGroup": "string"
      },
      "SecurityGroups": [ "string" ],
      "SharingSettings": {
         "NotebookOutputOption": "string",
         "S3KmsKeyId": "string",
         "S3OutputPath": "string"
      },
      "SpaceStorageSettings": {
         "DefaultEbsStorageSettings": {
            "DefaultEbsVolumeSizeInGb": number,
            "MaximumEbsVolumeSizeInGb": number
         }
      },
      "StudioWebPortal": "string",
      "StudioWebPortalSettings": {
         "ExecutionRoleSessionNameMode": "string",
         "HiddenAppTypes": [ "string" ],
         "HiddenInstanceTypes": [ "string" ],
         "HiddenMlTools": [ "string" ],
         "HiddenSageMakerImageVersionAliases": [
            {
               "SageMakerImageName": "string",
               "VersionAliases": [ "string" ]
            }
         ]
      },
      "TensorBoardAppSettings": {
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
```

```
            "TrainingPlanArn": "string"
         }
      }
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainId

The ID of the associated Domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

SingleSignOnUserIdentifier

A specifier for the type of value specified in SingleSignOnUserValue. Currently, the only supported value is "UserName". If the Domain's AuthMode is IAM Identity Center, this field is required. If the Domain's AuthMode is not IAM Identity Center, this field cannot be specified.

Type: String

Pattern: `UserName`

Required: No

SingleSignOnUserValue

The username of the associated AWS Single Sign-On User for this UserProfile. If the Domain's AuthMode is IAM Identity Center, this field is required, and must match a valid username of a user in your directory. If the Domain's AuthMode is not IAM Identity Center, this field cannot be specified.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

Tags

Each tag consists of a key and an optional value. Tag keys must be unique per resource.

Tags that you specify for the User Profile are also added to all Apps that the User Profile launches.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

UserProfileName

A name for the UserProfile. This value is not case sensitive.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

UserSettings

A collection of settings.

Type: UserSettings object

Required: No

Response Syntax

```
{
   "UserProfileArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

UserProfileArn

The user profile Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:user-profile/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateWorkforce

Service: Amazon SageMaker Service

Use this operation to create a workforce. This operation will return an error if a workforce already exists in the AWS Region that you specify. You can only create one workforce in each AWS Region per AWS account.

If you want to create a new workforce in an AWS Region where a workforce already exists, use the DeleteWorkforce API operation to delete the existing workforce and then use `CreateWorkforce` to create a new workforce.

To create a private workforce using Amazon Cognito, you must specify a Cognito user pool in `CognitoConfig`. You can also create an Amazon Cognito workforce using the Amazon SageMaker console. For more information, see Create a Private Workforce (Amazon Cognito).

To create a private workforce using your own OIDC Identity Provider (IdP), specify your IdP configuration in `OidcConfig`. Your OIDC IdP must support *groups* because groups are used by Ground Truth and Amazon A2I to create work teams. For more information, see Create a Private Workforce (OIDC IdP).

Request Syntax

```
{
   "CognitoConfig": {
      "ClientId": "string",
      "UserPool": "string"
   },
   "IpAddressType": "string",
   "OidcConfig": {
      "AuthenticationRequestExtraParams": {
         "string" : "string"
      },
      "AuthorizationEndpoint": "string",
      "ClientId": "string",
      "ClientSecret": "string",
      "Issuer": "string",
      "JwksUri": "string",
      "LogoutEndpoint": "string",
      "Scope": "string",
      "TokenEndpoint": "string",
      "UserInfoEndpoint": "string"
```

```
   },
   "SourceIpConfig": {
      "Cidrs": [ "string" ]
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "WorkforceName": "string",
   "WorkforceVpcConfig": {
      "SecurityGroupIds": [ "string" ],
      "Subnets": [ "string" ],
      "VpcId": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CognitoConfig

Use this parameter to configure an Amazon Cognito private workforce. A single Cognito workforce is created using and corresponds to a single Amazon Cognito user pool.

Do not use `OidcConfig` if you specify values for `CognitoConfig`.

Type: CognitoConfig object

Required: No

IpAddressType

Use this parameter to specify whether you want `IPv4` only or `dualstack` (`IPv4` and `IPv6`) to support your labeling workforce.

Type: String

Valid Values: `ipv4 | dualstack`

Required: No

OidcConfig

Use this parameter to configure a private workforce using your own OIDC Identity Provider.

Do not use `CognitoConfig` if you specify values for `OidcConfig`.

Type: OidcConfig object

Required: No

SourceIpConfig

A list of IP address ranges (CIDRs). Used to create an allow list of IP addresses for a private workforce. Workers will only be able to log in to their worker portal from an IP address within this range. By default, a workforce isn't restricted to specific IP addresses.

Type: SourceIpConfig object

Required: No

Tags

An array of key-value pairs that contain metadata to help you categorize and organize our workforce. Each tag consists of a key and a value, both of which you define.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

WorkforceName

The name of the private workforce.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9\-]){0,62}`

Required: Yes

WorkforceVpcConfig

Use this parameter to configure a workforce using VPC.

Type: WorkforceVpcConfigRequest object

Required: No

Response Syntax

```
{
   "WorkforceArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

WorkforceArn

The Amazon Resource Name (ARN) of the workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:workforce/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateWorkteam

Service: Amazon SageMaker Service

Creates a new work team for labeling your data. A work team is defined by one or more Amazon Cognito user pools. You must first create the user pools before you can create a work team.

You cannot create more than 25 work teams in an account and region.

Request Syntax

```
{
   "Description": "string",
   "MemberDefinitions": [
      {
         "CognitoMemberDefinition": {
            "ClientId": "string",
            "UserGroup": "string",
            "UserPool": "string"
         },
         "OidcMemberDefinition": {
            "Groups": [ "string" ]
         }
      }
   ],
   "NotificationConfiguration": {
      "NotificationTopicArn": "string"
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "WorkerAccessConfiguration": {
      "S3Presign": {
         "IamPolicyConstraints": {
            "SourceIp": "string",
            "VpcSourceIp": "string"
         }
      }
   },
   "WorkforceName": "string",
   "WorkteamName": "string"
```

```
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Description

A description of the work team.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `.+`

Required: Yes

MemberDefinitions

A list of `MemberDefinition` objects that contains objects that identify the workers that make up the work team.

Workforces can be created using Amazon Cognito or your own OIDC Identity Provider (IdP). For private workforces created using Amazon Cognito use `CognitoMemberDefinition`. For workforces created using your own OIDC identity provider (IdP) use `OidcMemberDefinition`. Do not provide input for both of these parameters in a single request.

For workforces created using Amazon Cognito, private work teams correspond to Amazon Cognito *user groups* within the user pool used to create a workforce. All of the `CognitoMemberDefinition` objects that make up the member definition must have the same `ClientId` and `UserPool` values. To add a Amazon Cognito user group to an existing worker pool, see Adding groups to a User Pool. For more information about user pools, see Amazon Cognito User Pools.

For workforces created using your own OIDC IdP, specify the user groups that you want to include in your private work team in `OidcMemberDefinition` by listing those groups in `Groups`.

Type: Array of MemberDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: Yes

NotificationConfiguration

Configures notification of workers regarding available or expiring work items.

Type: NotificationConfiguration object

Required: No

Tags

An array of key-value pairs.

For more information, see Resource Tag and Using Cost Allocation Tags in the *AWS Billing and Cost Management User Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

WorkerAccessConfiguration

Use this optional parameter to constrain access to an Amazon S3 resource based on the IP address using supported IAM global condition keys. The Amazon S3 resource is accessed in the worker portal using a Amazon S3 presigned URL.

Type: WorkerAccessConfiguration object

Required: No

WorkforceName

The name of the workforce.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9\-]){0,62}`

Required: No

WorkteamName

The name of the work team. Use this name to identify the work team.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "WorkteamArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

WorkteamArn

The Amazon Resource Name (ARN) of the work team. You can use this ARN to identify the work team.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:workteam/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteAction

Service: Amazon SageMaker Service

Deletes an action.

Request Syntax

```
{
   "ActionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ActionName

The name of the action to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "ActionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ActionArn

The Amazon Resource Name (ARN) of the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:action/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteAIBenchmarkJob

Service: Amazon SageMaker Service

Deletes the specified AI benchmark job.

Request Syntax

```
{
   "AIBenchmarkJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AIBenchmarkJobName

The name of the AI benchmark job to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AIBenchmarkJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIBenchmarkJobArn

The Amazon Resource Name (ARN) of the deleted benchmark job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-benchmark-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteAIRecommendationJob

Service: Amazon SageMaker Service

Deletes the specified AI recommendation job.

Request Syntax

```
{
   "AIRecommendationJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AIRecommendationJobName

The name of the AI recommendation job to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AIRecommendationJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIRecommendationJobArn

The Amazon Resource Name (ARN) of the deleted recommendation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-recommendation-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteAIWorkloadConfig

Service: Amazon SageMaker Service

Deletes the specified AI workload configuration. You cannot delete a configuration that is referenced by an active benchmark job.

Request Syntax

```
{
   "AIWorkloadConfigName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AIWorkloadConfigName

The name of the AI workload configuration to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AIWorkloadConfigArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIWorkloadConfigArn

The Amazon Resource Name (ARN) of the deleted AI workload configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-workload-config/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteAlgorithm

Service: Amazon SageMaker Service

Removes the specified algorithm from your account.

Request Syntax

```
{
   "AlgorithmName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AlgorithmName

The name of the algorithm to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteApp

Service: Amazon SageMaker Service

Used to stop and delete an app.

Request Syntax

```
{
   "AppName": "string",
   "AppType": "string",
   "DomainId": "string",
   "SpaceName": "string",
   "UserProfileName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AppName

The name of the app.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AppType

The type of app.

Type: String

Valid Values: `JupyterServer | KernelGateway | DetailedProfiler | TensorBoard | CodeEditor | JupyterLab | RStudioServerPro | RSessionGateway | Canvas`

Required: Yes

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

SpaceName

The name of the space. If this value is not set, then `UserProfileName` must be set.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

UserProfileName

The user profile name. If this value is not set, then `SpaceName` must be set.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteAppImageConfig

Service: Amazon SageMaker Service

Deletes an AppImageConfig.

Request Syntax

```
{
   "AppImageConfigName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AppImageConfigName

The name of the AppImageConfig to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteArtifact

Service: Amazon SageMaker Service

Deletes an artifact. Either `ArtifactArn` or `Source` must be specified.

Request Syntax

```
{
   "ArtifactArn": "string",
   "Source": {
      "SourceTypes": [
         {
            "SourceIdType": "string",
            "Value": "string"
         }
      ],
      "SourceUri": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ArtifactArn

The Amazon Resource Name (ARN) of the artifact to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:artifact/.*`

Required: No

Source

The URI of the source.

Type: ArtifactSource object

Required: No

Response Syntax

```
{
   "ArtifactArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ArtifactArn

The Amazon Resource Name (ARN) of the artifact.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:artifact/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteAssociation

Service: Amazon SageMaker Service

Deletes an association.

Request Syntax

```
{
   "DestinationArn": "string",
   "SourceArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DestinationArn

The Amazon Resource Name (ARN) of the destination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: Yes

SourceArn

The ARN of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: Yes

Response Syntax

```
{
   "DestinationArn": "string",
   "SourceArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

DestinationArn

The Amazon Resource Name (ARN) of the destination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

SourceArn

The ARN of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteCluster

Service: Amazon SageMaker Service

Delete a SageMaker HyperPod cluster.

Request Syntax

```
{
   "ClusterName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterName

The string name or the Amazon Resource Name (ARN) of the SageMaker HyperPod cluster to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

Response Syntax

```
{
   "ClusterArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ClusterArn

The Amazon Resource Name (ARN) of the SageMaker HyperPod cluster to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteClusterSchedulerConfig

Service: Amazon SageMaker Service

Deletes the cluster policy of the cluster.

Request Syntax

```
{
   "ClusterSchedulerConfigId": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterSchedulerConfigId

ID of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[a-z0-9]{12}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteCodeRepository

Service: Amazon SageMaker Service

Deletes the specified Git repository from your account.

Request Syntax

```
{
   "CodeRepositoryName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CodeRepositoryName

The name of the Git repository to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteCompilationJob

Service: Amazon SageMaker Service

Deletes the specified compilation job. This action deletes only the compilation job resource in Amazon SageMaker AI. It doesn't delete other resources that are related to that job, such as the model artifacts that the job creates, the compilation logs in CloudWatch, the compiled model, or the IAM role.

You can delete a compilation job only if its current status is `COMPLETED`, `FAILED`, or `STOPPED`. If the job status is `STARTING` or `INPROGRESS`, stop the job, and then delete it after its status becomes `STOPPED`.

Request Syntax

```
{
   "CompilationJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CompilationJobName

The name of the compilation job to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteComputeQuota

Service: Amazon SageMaker Service

Deletes the compute allocation from the cluster.

Request Syntax

```
{
   "ComputeQuotaId": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ComputeQuotaId

ID of the compute allocation definition.

Type: String

Pattern: `[a-z0-9]{12}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteContext

Service: Amazon SageMaker Service

Deletes an context.

Request Syntax

```
{
   "ContextName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ContextName

The name of the context to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "ContextArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ContextArn

The Amazon Resource Name (ARN) of the context.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:context/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteDataQualityJobDefinition

Service: Amazon SageMaker Service

Deletes a data quality monitoring job definition.

Request Syntax

```
{
   "JobDefinitionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobDefinitionName

The name of the data quality monitoring job definition to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteDeviceFleet

Service: Amazon SageMaker Service

Deletes a fleet.

Request Syntax

```
{
   "DeviceFleetName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeviceFleetName

The name of the fleet to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteDomain

Service: Amazon SageMaker Service

Used to delete a domain. If you onboarded with IAM mode, you will need to delete your domain to onboard again using IAM Identity Center. Use with caution. All of the members of the domain will lose access to their EFS volume, including data, notebooks, and other artifacts.

Request Syntax

```
{
   "DomainId": "string",
   "RetentionPolicy": {
      "HomeEfsFileSystem": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

RetentionPolicy

The retention policy for this domain, which specifies whether resources will be retained after the Domain is deleted. By default, all resources are retained (not automatically deleted).

Type: RetentionPolicy object

Required: No

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteEdgeDeploymentPlan

Service: Amazon SageMaker Service

Deletes an edge deployment plan if (and only if) all the stages in the plan are inactive or there are no stages in the plan.

Request Syntax

```
{
   "EdgeDeploymentPlanName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EdgeDeploymentPlanName

The name of the edge deployment plan to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteEdgeDeploymentStage

Service: Amazon SageMaker Service

Delete a stage in an edge deployment plan if (and only if) the stage is inactive.

Request Syntax

```
{
   "EdgeDeploymentPlanName": "string",
   "StageName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EdgeDeploymentPlanName

The name of the edge deployment plan from which the stage will be deleted.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

StageName

The name of the stage.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteEndpoint

Service: Amazon SageMaker Service

Deletes an endpoint. SageMaker frees up all of the resources that were deployed when the endpoint was created.

SageMaker retires any custom KMS key grants associated with the endpoint, meaning you don't need to use the RevokeGrant API call.

When you delete your endpoint, SageMaker asynchronously deletes associated endpoint resources such as KMS key grants. You might still see these resources in your account for a few minutes after deleting your endpoint. Do not delete or revoke the permissions for your `ExecutionRoleArn`, otherwise SageMaker cannot delete these resources.

Request Syntax

```
{
   "EndpointName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EndpointName

The name of the endpoint that you want to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteEndpointConfig

Service: Amazon SageMaker Service

Deletes an endpoint configuration. The `DeleteEndpointConfig` API deletes only the specified configuration. It does not delete endpoints created using the configuration.

You must not delete an `EndpointConfig` in use by an endpoint that is live or while the `UpdateEndpoint` or `CreateEndpoint` operations are being performed on the endpoint. If you delete the `EndpointConfig` of an endpoint that is active or being created or updated you may lose visibility into the instance type the endpoint is using. The endpoint must be deleted in order to stop incurring charges.

Request Syntax

```
{
   "EndpointConfigName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EndpointConfigName

The name of the endpoint configuration that you want to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteExperiment

Service: Amazon SageMaker Service

Deletes an SageMaker experiment. All trials associated with the experiment must be deleted first. Use the ListTrials API to get a list of the trials associated with the experiment.

Request Syntax

```
{
   "ExperimentName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ExperimentName

The name of the experiment to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "ExperimentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ExperimentArn

The Amazon Resource Name (ARN) of the experiment that is being deleted.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteFeatureGroup

Service: Amazon SageMaker Service

Delete the `FeatureGroup` and any data that was written to the `OnlineStore` of the `FeatureGroup`. Data cannot be accessed from the `OnlineStore` immediately after `DeleteFeatureGroup` is called.

Data written into the `OfflineStore` will not be deleted. The AWS Glue database and tables that are automatically created for your `OfflineStore` are not deleted.

Note that it can take approximately 10-15 minutes to delete an `OnlineStore` `FeatureGroup` with the `InMemory` `StorageType`.

Request Syntax

```
{
   "FeatureGroupName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

FeatureGroupName

The name of the `FeatureGroup` you want to delete. The name must be unique within an AWS Region in an AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteFlowDefinition

Service: Amazon SageMaker Service

Deletes the specified flow definition.

Request Syntax

```
{
   "FlowDefinitionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

FlowDefinitionName

The name of the flow definition you are deleting.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteHub

Service: Amazon SageMaker Service

Delete a hub.

Request Syntax

```
{
   "HubName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HubName

The name of the hub to delete.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteHubContent

Service: Amazon SageMaker Service

Delete the contents of a hub.

Request Syntax

```
{
   "HubContentName": "string",
   "HubContentType": "string",
   "HubContentVersion": "string",
   "HubName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HubContentName

The name of the content that you want to delete from a hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubContentType

The type of content that you want to delete from a hub.

Type: String

Valid Values: `Model | Notebook | ModelReference | DataSet | JsonDoc`

Required: Yes

HubContentVersion

The version of the content that you want to delete from a hub.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: Yes

HubName

The name of the hub that you want to delete content in.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteHubContentReference

Service: Amazon SageMaker Service

Delete a hub content reference in order to remove a model from a private hub.

Request Syntax

```
{
   "HubContentName": "string",
   "HubContentType": "string",
   "HubName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HubContentName

The name of the hub content to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubContentType

The type of hub content reference to delete. The only supported type of hub content reference to delete is `ModelReference`.

Type: String

Valid Values: `Model | Notebook | ModelReference | DataSet | JsonDoc`

Required: Yes

HubName

The name of the hub to delete the hub content reference from.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteHumanTaskUi

Service: Amazon SageMaker Service

Use this operation to delete a human task user interface (worker task template).

To see a list of human task user interfaces (work task templates) in your account, use ListHumanTaskUis. When you delete a worker task template, it no longer appears when you call `ListHumanTaskUis`.

Request Syntax

```
{
   "HumanTaskUiName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HumanTaskUiName

The name of the human task user interface (work task template) you want to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9])*`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteHyperParameterTuningJob

Service: Amazon SageMaker Service

Deletes a hyperparameter tuning job. The `DeleteHyperParameterTuningJob` API deletes only the tuning job entry that was created in SageMaker when you called the `CreateHyperParameterTuningJob` API. It does not delete training jobs, artifacts, or the IAM role that you specified when creating the model.

Request Syntax

```
{
   "HyperParameterTuningJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HyperParameterTuningJobName

The name of the hyperparameter tuning job that you want to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteImage

Service: Amazon SageMaker Service

Deletes a SageMaker AI image and all versions of the image. The container images aren't deleted.

Request Syntax

```
{
   "ImageName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ImageName

The name of the image to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteImageVersion

Service: Amazon SageMaker Service

Deletes a version of a SageMaker AI image. The container image the version represents isn't deleted.

Request Syntax

```
{
   "Alias": "string",
   "ImageName": "string",
   "Version": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Alias

The alias of the image to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(?!^[.-])^([a-zA-Z0-9-_.]+)`

Required: No

ImageName

The name of the image to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

Version

The version to delete.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteInferenceComponent

Service: Amazon SageMaker Service

Deletes an inference component.

Request Syntax

```
{
   "InferenceComponentName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

InferenceComponentName

The name of the inference component to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteInferenceExperiment

Service: Amazon SageMaker Service

Deletes an inference experiment.

> **Note**
>
> This operation does not delete your endpoint, variants, or any underlying resources. This operation only deletes the metadata of your experiment.

Request Syntax

```
{
   "Name": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Name

The name of the inference experiment you want to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "InferenceExperimentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

InferenceExperimentArn

The ARN of the deleted inference experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:inference-experiment/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteJob

Service: Amazon SageMaker Service

Deletes a job. This operation is idempotent. If the job is currently running, you must stop it before deleting it by calling `StopJob`.

The following operations are related to `DeleteJob`:

- `CreateJob`
- `StopJob`
- `DescribeJob`

Request Syntax

```
{
   "JobCategory": "string",
   "JobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobCategory

The category of the job to delete.

Type: String

Valid Values: `AgentRFT | AgentRFTEvaluation`

Required: Yes

JobName

The name of the job to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteMlflowApp

Service: Amazon SageMaker Service

Deletes an MLflow App.

Request Syntax

```
{
   "Arn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Arn

The ARN of the MLflow App to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-app/.*`

Required: Yes

Response Syntax

```
{
   "Arn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The ARN of the deleted MLflow App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-app/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteMlflowTrackingServer

Service: Amazon SageMaker Service

Deletes an MLflow Tracking Server. For more information, see Clean up MLflow resources.

Request Syntax

```
{
   "TrackingServerName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrackingServerName

The name of the the tracking server to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: Yes

Response Syntax

```
{
   "TrackingServerArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrackingServerArn

A `TrackingServerArn` object, the ARN of the tracking server that is deleted if successfully found.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-tracking-server/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteModel

Service: Amazon SageMaker Service

Deletes a model. The `DeleteModel` API deletes only the model entry that was created in SageMaker when you called the `CreateModel` API. It does not delete model artifacts, inference code, or the IAM role that you specified when creating the model.

Request Syntax

```
{
   "ModelName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelName

The name of the model to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteModelBiasJobDefinition

Service: Amazon SageMaker Service

Deletes an Amazon SageMaker AI model bias job definition.

Request Syntax

```
{
   "JobDefinitionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobDefinitionName

The name of the model bias job definition to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteModelCard

Service: Amazon SageMaker Service

Deletes an Amazon SageMaker Model Card.

Request Syntax

```
{
   "ModelCardName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelCardName

The name of the model card to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteModelExplainabilityJobDefinition

Service: Amazon SageMaker Service

Deletes an Amazon SageMaker AI model explainability job definition.

Request Syntax

```
{
   "JobDefinitionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobDefinitionName

The name of the model explainability job definition to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteModelPackage

Service: Amazon SageMaker Service

Deletes a model package.

A model package is used to create SageMaker models or list on AWS Marketplace. Buyers can subscribe to model packages listed on AWS Marketplace to create models in SageMaker.

Request Syntax

```
{
   "ModelPackageName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelPackageName

The name or Amazon Resource Name (ARN) of the model package to delete.

When you specify a name, the name must have 1 to 63 characters. Valid characters are a-z, A-Z, 0-9, and - (hyphen).

Type: String

Length Constraints: Minimum length of 1. Maximum length of 176.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*\/)?([a-zA-Z0-9]([a-zA-Z0-9-]){0,62})(?<!-)(\/[0-9]{1,9})?`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteModelPackageGroup

Service: Amazon SageMaker Service

Deletes the specified model group.

Request Syntax

```
{
   "ModelPackageGroupName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelPackageGroupName

The name of the model group to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 170.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*\/)?([a-zA-Z0-9]([a-zA-Z0-9-]){0,62})(?<!-)`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteModelPackageGroupPolicy

Service: Amazon SageMaker Service

Deletes a model group resource policy.

Request Syntax

```
{
   "ModelPackageGroupName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelPackageGroupName

The name of the model group for which to delete the policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteModelQualityJobDefinition

Service: Amazon SageMaker Service

Deletes the secified model quality monitoring job definition.

Request Syntax

```
{
   "JobDefinitionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobDefinitionName

The name of the model quality monitoring job definition to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteMonitoringSchedule

Service: Amazon SageMaker Service

Deletes a monitoring schedule. Also stops the schedule had not already been stopped. This does not delete the job execution history of the monitoring schedule.

Request Syntax

```
{
   "MonitoringScheduleName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MonitoringScheduleName

The name of the monitoring schedule to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteNotebookInstance

Service: Amazon SageMaker Service

Deletes an SageMaker AI notebook instance. Before you can delete a notebook instance, you must call the `StopNotebookInstance` API.

> **Important**
>
> When you delete a notebook instance, you lose all of your data. SageMaker AI removes the ML compute instance, and deletes the ML storage volume and the network interface associated with the notebook instance.

Request Syntax

```
{
   "NotebookInstanceName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

NotebookInstanceName

The name of the SageMaker AI notebook instance to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteNotebookInstanceLifecycleConfig

Service: Amazon SageMaker Service

Deletes a notebook instance lifecycle configuration.

Request Syntax

```
{
   "NotebookInstanceLifecycleConfigName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

NotebookInstanceLifecycleConfigName

The name of the lifecycle configuration to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteOptimizationJob

Service: Amazon SageMaker Service

Deletes an optimization job.

Request Syntax

```
{
   "OptimizationJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

OptimizationJobName

The name that you assigned to the optimization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeletePartnerApp

Service: Amazon SageMaker Service

Deletes a SageMaker Partner AI App.

Request Syntax

```
{
   "Arn": "string",
   "ClientToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Arn

The ARN of the SageMaker Partner AI App to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:partner-app\/app-[A-Z0-9]{12}`

Required: Yes

ClientToken

A unique token that guarantees that the call to this API is idempotent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 36.

Pattern: `[a-zA-Z0-9-]+`

Required: No

Response Syntax

```
{
   "Arn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The ARN of the SageMaker Partner AI App that was deleted.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:partner-app\/app-[A-Z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeletePipeline

Service: Amazon SageMaker Service

Deletes a pipeline if there are no running instances of the pipeline. To delete a pipeline, you must stop all running instances of the pipeline using the `StopPipelineExecution` API. When you delete a pipeline, all instances of the pipeline are deleted.

Request Syntax

```
{
   "ClientRequestToken": "string",
   "PipelineName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClientRequestToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the operation. An idempotent operation completes no more than one time.

Type: String

Length Constraints: Minimum length of 32. Maximum length of 128.

Required: Yes

PipelineName

The name of the pipeline to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: Yes

Response Syntax

```
{
   "PipelineArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

PipelineArn

The Amazon Resource Name (ARN) of the pipeline to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteProcessingJob

Service: Amazon SageMaker Service

Deletes a processing job. After Amazon SageMaker deletes a processing job, all of the metadata for the processing job is lost. You can delete only processing jobs that are in a terminal state (`Stopped`, `Failed`, or `Completed`). You cannot delete a job that is in the `InProgress` or `Stopping` state. After deleting the job, you can reuse its name to create another processing job.

Request Syntax

```
{
   "ProcessingJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ProcessingJobName

The name of the processing job to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteProject

Service: Amazon SageMaker Service

Delete the specified project.

Request Syntax

```
{
   "ProjectName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ProjectName

The name of the project to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteSpace

Service: Amazon SageMaker Service

Used to delete a space.

Request Syntax

```
{
   "DomainId": "string",
   "SpaceName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainId

The ID of the associated domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

SpaceName

The name of the space.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteStudioLifecycleConfig

Service: Amazon SageMaker Service

Deletes the Amazon SageMaker AI Studio Lifecycle Configuration. In order to delete the Lifecycle Configuration, there must be no running apps using the Lifecycle Configuration. You must also remove the Lifecycle Configuration from UserSettings in all Domains and UserProfiles.

Request Syntax

```
{
   "StudioLifecycleConfigName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

StudioLifecycleConfigName

The name of the Amazon SageMaker AI Studio Lifecycle Configuration to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteTags

Service: Amazon SageMaker Service

Deletes the specified tags from an SageMaker resource.

To list a resource's tags, use the `ListTags` API.

> **Note**
>
> When you call this API to delete tags from a hyperparameter tuning job, the deleted tags are not removed from training jobs that the hyperparameter tuning job launched before you called this API.

> **Note**
>
> When you call this API to delete tags from a SageMaker Domain or User Profile, the deleted tags are not removed from Apps that the SageMaker Domain or User Profile launched before you called this API.

Request Syntax

```
{
   "ResourceArn": "string",
   "TagKeys": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ResourceArn

The Amazon Resource Name (ARN) of the resource whose tags you want to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z-]*:sagemaker:[a-z0-9-]*:[0-9]{12}:.+`

Required: Yes

TagKeys

An array or one or more tag keys to delete.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 50 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `([\p{L}\p{Z}\p{N}_.:/=+\-@]*)`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteTrainingJob

Service: Amazon SageMaker Service

Deletes a training job. After SageMaker deletes a training job, all of the metadata for the training job is lost. You can delete only training jobs that are in a terminal state (`Stopped`, `Failed`, or `Completed`) and don't retain an `Available` managed warm pool. You cannot delete a job that is in the `InProgress` or `Stopping` state. After deleting the job, you can reuse its name to create another training job.

Request Syntax

```
{
   "TrainingJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrainingJobName

The name of the training job to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteTrial

Service: Amazon SageMaker Service

Deletes the specified trial. All trial components that make up the trial must be deleted first. Use the DescribeTrialComponent API to get the list of trial components.

Request Syntax

```
{
   "TrialName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrialName

The name of the trial to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "TrialArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrialArn

The Amazon Resource Name (ARN) of the trial that is being deleted.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteTrialComponent

Service: Amazon SageMaker Service

Deletes the specified trial component. A trial component must be disassociated from all trials before the trial component can be deleted. To disassociate a trial component from a trial, call the DisassociateTrialComponent API.

Request Syntax

```
{
   "TrialComponentName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrialComponentName

The name of the component to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "TrialComponentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrialComponentArn

The Amazon Resource Name (ARN) of the component is being deleted.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial-component/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteUserProfile

Service: Amazon SageMaker Service

Deletes a user profile. When a user profile is deleted, the user loses access to their EFS volume, including data, notebooks, and other artifacts.

Request Syntax

```
{
   "DomainId": "string",
   "UserProfileName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

UserProfileName

The user profile name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteWorkforce

Service: Amazon SageMaker Service

Use this operation to delete a workforce.

If you want to create a new workforce in an AWS Region where a workforce already exists, use this operation to delete the existing workforce and then use CreateWorkforce to create a new workforce.

> **Important**
>
> If a private workforce contains one or more work teams, you must use the DeleteWorkteam operation to delete all work teams before you delete the workforce. If you try to delete a workforce that contains one or more work teams, you will receive a `ResourceInUse` error.

Request Syntax

```
{
   "WorkforceName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

WorkforceName

The name of the workforce.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9\-]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteWorkteam

Service: Amazon SageMaker Service

Deletes an existing work team. This operation can't be undone.

Request Syntax

```
{
   "WorkteamName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

WorkteamName

The name of the work team to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "Success": boolean
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Success

Returns `true` if the work team was successfully deleted; otherwise, returns `false`.

Type: Boolean

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeregisterDevices

Service: Amazon SageMaker Service

Deregisters the specified devices. After you deregister a device, you will need to re-register the devices.

Request Syntax

```
{
   "DeviceFleetName": "string",
   "DeviceNames": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeviceFleetName

The name of the fleet the devices belong to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

DeviceNames

The unique IDs of the devices.

Type: Array of strings

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeAction

Service: Amazon SageMaker Service

Describes an action.

Request Syntax

```
{
   "ActionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ActionName

The name of the action to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial|experiment-trial-component|artifact|action|context)\/)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,119})`

Required: Yes

Response Syntax

```
{
   "ActionArn": "string",
   "ActionName": "string",
   "ActionType": "string",
   "CreatedBy": {
      "DomainId": "string",
```

```
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "CreationTime": number,
   "Description": "string",
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "LineageGroupArn": "string",
   "MetadataProperties": {
      "CommitId": "string",
      "GeneratedBy": "string",
      "ProjectId": "string",
      "Repository": "string"
   },
   "Properties": {
      "string" : "string"
   },
   "Source": {
      "SourceId": "string",
      "SourceType": "string",
      "SourceUri": "string"
   },
   "Status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ActionArn

The Amazon Resource Name (ARN) of the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:action/.*`

ActionName

The name of the action.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial|experiment-trial-component|artifact|action|context)\/)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,119})`

ActionType

The type of the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

When the action was created.

Type: Timestamp

Description

The description of the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

When the action was last modified.

Type: Timestamp

LineageGroupArn

The Amazon Resource Name (ARN) of the lineage group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:lineage-group/.*`

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Properties

A list of the action's properties.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 30 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2500.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Source

The source of the action.

Type: ActionSource object

Status

The status of the action.

Type: String

Valid Values: `Unknown | InProgress | Completed | Failed | Stopping | Stopped`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeAIBenchmarkJob

Service: Amazon SageMaker Service

Returns details of an AI benchmark job, including its status, configuration, target endpoint, and timing information.

Request Syntax

```
{
   "AIBenchmarkJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AIBenchmarkJobName

The name of the AI benchmark job to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AIBenchmarkJobArn": "string",
   "AIBenchmarkJobName": "string",
   "AIBenchmarkJobStatus": "string",
   "AIWorkloadConfigIdentifier": "string",
   "BenchmarkTarget": { ... },
   "CreationTime": number,
   "EndTime": number,
```

```
   "FailureReason": "string",
   "NetworkConfig": {
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
   },
   "OutputConfig": {
      "CloudWatchLogs": [
         {
            "LogGroupArn": "string",
            "LogStreamName": "string"
         }
      ],
      "MlflowConfig": {
         "MlflowExperimentName": "string",
         "MlflowResourceArn": "string",
         "MlflowRunName": "string"
      },
      "S3OutputLocation": "string"
   },
   "RoleArn": "string",
   "StartTime": number,
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIBenchmarkJobArn

The Amazon Resource Name (ARN) of the AI benchmark job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-benchmark-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

AIBenchmarkJobName

The name of the AI benchmark job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

AIBenchmarkJobStatus

The status of the AI benchmark job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

AIWorkloadConfigIdentifier

The name or Amazon Resource Name (ARN) of the AI workload configuration used for this benchmark job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*/)?([a-zA-Z0-9]([a-zA-Z0-9\-]){0,62})(?<!-)`

BenchmarkTarget

The target endpoint that was benchmarked.

Type: AIBenchmarkTarget object

Note: This object is a Union. Only one member of this object can be specified or returned.

CreationTime

A timestamp that indicates when the benchmark job was created.

Type: Timestamp

EndTime

A timestamp that indicates when the benchmark job completed.

Type: Timestamp

FailureReason

If the benchmark job failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

NetworkConfig

The network configuration for the benchmark job.

Type: AIBenchmarkNetworkConfig object

OutputConfig

The output configuration for the benchmark job, including the Amazon S3 output location and CloudWatch log information.

Type: AIBenchmarkOutputResult object

RoleArn

The Amazon Resource Name (ARN) of the IAM role used by the benchmark job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

StartTime

A timestamp that indicates when the benchmark job started running.

Type: Timestamp

Tags

The tags associated with the benchmark job.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeAIRecommendationJob

Service: Amazon SageMaker Service

Returns details of an AI recommendation job, including its status, model source, performance targets, optimization recommendations, and deployment configurations.

Request Syntax

```
{
   "AIRecommendationJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AIRecommendationJobName

The name of the AI recommendation job to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AIRecommendationJobArn": "string",
   "AIRecommendationJobName": "string",
   "AIRecommendationJobStatus": "string",
   "AIWorkloadConfigIdentifier": "string",
   "ComputeSpec": {
      "CapacityReservationConfig": {
         "CapacityReservationPreference": "string",
         "MlReservationArns": [ "string" ]
      },
```

```
      "InstanceTypes": [ "string" ]
   },
   "CreationTime": number,
   "EndTime": number,
   "FailureReason": "string",
   "InferenceSpecification": {
      "Framework": "string"
   },
   "ModelSource": { ... },
   "OptimizeModel": boolean,
   "OutputConfig": {
      "MlflowConfig": {
         "MlflowExperimentName": "string",
         "MlflowResourceArn": "string",
         "MlflowRunName": "string"
      },
      "ModelPackageGroupIdentifier": "string",
      "S3OutputLocation": "string"
   },
   "PerformanceTarget": {
      "Constraints": [
         {
            "Metric": "string"
         }
      ]
   },
   "Recommendations": [
      {
         "AIBenchmarkJobArn": "string",
         "DeploymentConfiguration": {
            "CopyCountPerInstance": number,
            "EnvironmentVariables": {
               "string" : "string"
            },
            "ImageUri": "string",
            "InstanceCount": number,
            "InstanceType": "string",
            "S3": [
               {
                  "ChannelName": "string",
                  "Uri": "string"
               }
            ]
         },
```

```
         "ExpectedPerformance": [
            {
               "Metric": "string",
               "Stat": "string",
               "Unit": "string",
               "Value": "string"
            }
         ],
         "ModelDetails": {
            "InferenceSpecificationName": "string",
            "InstanceDetails": [
               {
                  "CopyCountPerInstance": number,
                  "InstanceCount": number,
                  "InstanceType": "string"
               }
            ],
            "ModelPackageArn": "string"
         },
         "OptimizationDetails": [
            {
               "OptimizationConfig": {
                  "string" : "string"
               },
               "OptimizationType": "string"
            }
         ],
         "RecommendationDescription": "string"
      }
   ],
   "RoleArn": "string",
   "StartTime": number,
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIRecommendationJobArn

The Amazon Resource Name (ARN) of the AI recommendation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-recommendation-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

AIRecommendationJobName

The name of the AI recommendation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

AIRecommendationJobStatus

The status of the AI recommendation job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

AIWorkloadConfigIdentifier

The name or Amazon Resource Name (ARN) of the AI workload configuration used for this recommendation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*/)?([a-zA-Z0-9]([a-zA-Z0-9\-]){0,62})(?<!-)`

ComputeSpec

The compute resource specification for the recommendation job.

Type: AIRecommendationComputeSpec object

CreationTime

A timestamp that indicates when the recommendation job was created.

Type: Timestamp

EndTime

A timestamp that indicates when the recommendation job completed.

Type: Timestamp

FailureReason

If the recommendation job failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

InferenceSpecification

The inference framework configuration.

Type: AIRecommendationInferenceSpecification object

ModelSource

The source of the model that was analyzed.

Type: AIModelSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

OptimizeModel

Whether model optimization techniques were allowed.

Type: Boolean

OutputConfig

The output configuration for the recommendation job.

Type: AIRecommendationOutputResult object

PerformanceTarget

The performance targets specified for the recommendation job.

Type: AIRecommendationPerformanceTarget object

Recommendations

The list of optimization recommendations generated by the job. Each recommendation includes optimization details, deployment configuration, expected performance metrics, and the associated benchmark job ARN.

Type: Array of AIRecommendation objects

RoleArn

The Amazon Resource Name (ARN) of the IAM role used by the recommendation job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

StartTime

A timestamp that indicates when the recommendation job started running.

Type: Timestamp

Tags

The tags associated with the recommendation job.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeAIWorkloadConfig

Service: Amazon SageMaker Service

Returns details of an AI workload configuration, including the dataset configuration, benchmark tool settings, tags, and creation time.

Request Syntax

```
{
   "AIWorkloadConfigName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AIWorkloadConfigName

The name of the AI workload configuration to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AIWorkloadConfigArn": "string",
   "AIWorkloadConfigName": "string",
   "AIWorkloadConfigs": {
      "WorkloadSpec": { ... }
   },
   "CreationTime": number,
   "DatasetConfig": { ... },
   "Tags": [
```

```
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIWorkloadConfigArn

The Amazon Resource Name (ARN) of the AI workload configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-workload-config/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

AIWorkloadConfigName

The name of the AI workload configuration.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

AIWorkloadConfigs

The benchmark tool configuration and workload specification.

Type: AIWorkloadConfigs object

CreationTime

A timestamp that indicates when the AI workload configuration was created.

Type: Timestamp

DatasetConfig

The dataset configuration for the workload.

Type: AIDatasetConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Tags

The tags associated with the AI workload configuration.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeAlgorithm

Service: Amazon SageMaker Service

Returns a description of the specified algorithm that is in your account.

Request Syntax

```
{
   "AlgorithmName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AlgorithmName

The name of the algorithm to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 170.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*\/)?([a-zA-Z0-9]([a-zA-Z0-9-]){0,62})(?<!-)`

Required: Yes

Response Syntax

```
{
   "AlgorithmArn": "string",
   "AlgorithmDescription": "string",
   "AlgorithmName": "string",
   "AlgorithmStatus": "string",
   "AlgorithmStatusDetails": {
      "ImageScanStatuses": [
         {
            "FailureReason": "string",
            "Name": "string",
```

```
            "Status": "string"
         }
      ],
      "ValidationStatuses": [
         {
            "FailureReason": "string",
            "Name": "string",
            "Status": "string"
         }
      ]
   },
   "CertifyForMarketplace": boolean,
   "CreationTime": number,
   "InferenceSpecification": {
      "Containers": [
         {
            "AdditionalModelDataSources": [
               {
                  "ChannelName": "string",
                  "S3DataSource": {
                     "CompressionType": "string",
                     "ETag": "string",
                     "HubAccessConfig": {
                        "HubContentArn": "string"
                     },
                     "ManifestEtag": "string",
                     "ManifestS3Uri": "string",
                     "ModelAccessConfig": {
                        "AcceptEula": boolean
                     },
                     "S3DataType": "string",
                     "S3Uri": "string"
                  }
               }
            ],
            "AdditionalS3DataSource": {
               "CompressionType": "string",
               "ETag": "string",
               "S3DataType": "string",
               "S3Uri": "string"
            },
            "BaseModel": {
               "HubContentName": "string",
               "HubContentVersion": "string",
```

```
            "RecipeName": "string"
         },
         "ContainerHostname": "string",
         "Environment": {
            "string" : "string"
         },
         "Framework": "string",
         "FrameworkVersion": "string",
         "Image": "string",
         "ImageDigest": "string",
         "IsCheckpoint": boolean,
         "ModelDataETag": "string",
         "ModelDataSource": {
            "S3DataSource": {
               "CompressionType": "string",
               "ETag": "string",
               "HubAccessConfig": {
                  "HubContentArn": "string"
               },
               "ManifestEtag": "string",
               "ManifestS3Uri": "string",
               "ModelAccessConfig": {
                  "AcceptEula": boolean
               },
               "S3DataType": "string",
               "S3Uri": "string"
            }
         },
         "ModelDataUrl": "string",
         "ModelInput": {
            "DataInputConfig": "string"
         },
         "NearestModelName": "string",
         "ProductId": "string"
      }
   ],
   "SupportedContentTypes": [ "string" ],
   "SupportedRealtimeInferenceInstanceTypes": [ "string" ],
   "SupportedResponseMIMETypes": [ "string" ],
   "SupportedTransformInstanceTypes": [ "string" ]
},
"ProductId": "string",
"TrainingSpecification": {
   "AdditionalS3DataSource": {
```

```
            "CompressionType": "string",
            "ETag": "string",
            "S3DataType": "string",
            "S3Uri": "string"
         },
         "MetricDefinitions": [
            {
               "Name": "string",
               "Regex": "string"
            }
         ],
         "SupportedHyperParameters": [
            {
               "DefaultValue": "string",
               "Description": "string",
               "IsRequired": boolean,
               "IsTunable": boolean,
               "Name": "string",
               "Range": {
                  "CategoricalParameterRangeSpecification": {
                     "Values": [ "string" ]
                  },
                  "ContinuousParameterRangeSpecification": {
                     "MaxValue": "string",
                     "MinValue": "string"
                  },
                  "IntegerParameterRangeSpecification": {
                     "MaxValue": "string",
                     "MinValue": "string"
                  }
               },
               "Type": "string"
            }
         ],
         "SupportedTrainingInstanceTypes": [ "string" ],
         "SupportedTuningJobObjectiveMetrics": [
            {
               "MetricName": "string",
               "Type": "string"
            }
         ],
         "SupportsDistributedTraining": boolean,
         "TrainingChannels": [
            {
```

```
            "Description": "string",
            "IsRequired": boolean,
            "Name": "string",
            "SupportedCompressionTypes": [ "string" ],
            "SupportedContentTypes": [ "string" ],
            "SupportedInputModes": [ "string" ]
         }
      ],
      "TrainingImage": "string",
      "TrainingImageDigest": "string"
   },
   "ValidationSpecification": {
      "ValidationProfiles": [
         {
            "ProfileName": "string",
            "TrainingJobDefinition": {
               "HyperParameters": {
                  "string" : "string"
               },
               "InputDataConfig": [
                  {
                     "ChannelName": "string",
                     "CompressionType": "string",
                     "ContentType": "string",
                     "DataSource": {
                        "DatasetSource": {
                           "DatasetArn": "string"
                        },
                        "FileSystemDataSource": {
                           "DirectoryPath": "string",
                           "FileSystemAccessMode": "string",
                           "FileSystemId": "string",
                           "FileSystemType": "string"
                        },
                        "S3DataSource": {
                           "AttributeNames": [ "string" ],
                           "HubAccessConfig": {
                              "HubContentArn": "string"
                           },
                           "InstanceGroupNames": [ "string" ],
                           "ModelAccessConfig": {
                              "AcceptEula": boolean
                           },
                           "S3DataDistributionType": "string",
```

```
                        "S3DataType": "string",
                        "S3Uri": "string"
                     }
                  },
                  "InputMode": "string",
                  "RecordWrapperType": "string",
                  "ShuffleConfig": {
                     "Seed": number
                  }
               }
            ],
            "OutputDataConfig": {
               "CompressionType": "string",
               "KmsKeyId": "string",
               "S3OutputPath": "string"
            },
            "ResourceConfig": {
               "InstanceCount": number,
               "InstanceGroups": [
                  {
                     "InstanceCount": number,
                     "InstanceGroupName": "string",
                     "InstanceType": "string"
                  }
               ],
               "InstancePlacementConfig": {
                  "EnableMultipleJobs": boolean,
                  "PlacementSpecifications": [
                     {
                        "InstanceCount": number,
                        "UltraServerId": "string"
                     }
                  ]
               },
               "InstanceType": "string",
               "KeepAlivePeriodInSeconds": number,
               "TrainingPlanArn": "string",
               "VolumeKmsKeyId": "string",
               "VolumeSizeInGB": number
            },
            "StoppingCondition": {
               "MaxPendingTimeInSeconds": number,
               "MaxRuntimeInSeconds": number,
               "MaxWaitTimeInSeconds": number
```

```
            },
            "TrainingInputMode": "string"
         },
         "TransformJobDefinition": {
            "BatchStrategy": "string",
            "Environment": {
               "string" : "string"
            },
            "MaxConcurrentTransforms": number,
            "MaxPayloadInMB": number,
            "TransformInput": {
               "CompressionType": "string",
               "ContentType": "string",
               "DataSource": {
                  "S3DataSource": {
                     "S3DataType": "string",
                     "S3Uri": "string"
                  }
               },
               "SplitType": "string"
            },
            "TransformOutput": {
               "Accept": "string",
               "AssembleWith": "string",
               "KmsKeyId": "string",
               "S3OutputPath": "string"
            },
            "TransformResources": {
               "InstanceCount": number,
               "InstanceType": "string",
               "TransformAmiVersion": "string",
               "VolumeKmsKeyId": "string"
            }
         }
      },
      "ValidationRole": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AlgorithmArn

The Amazon Resource Name (ARN) of the algorithm.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:algorithm/[\S]{1,2048}`

AlgorithmDescription

A brief summary about the algorithm.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

AlgorithmName

The name of the algorithm being described.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

AlgorithmStatus

The current status of the algorithm.

Type: String

Valid Values: `Pending | InProgress | Completed | Failed | Deleting`

AlgorithmStatusDetails

Details about the current status of the algorithm.

Type: AlgorithmStatusDetails object

CertifyForMarketplace

Whether the algorithm is certified to be listed in AWS Marketplace.

Type: Boolean

CreationTime

A timestamp specifying when the algorithm was created.

Type: Timestamp

InferenceSpecification

Details about inference jobs that the algorithm runs.

Type: InferenceSpecification object

ProductId

The product identifier of the algorithm.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

TrainingSpecification

Details about training jobs run by this algorithm.

Type: TrainingSpecification object

ValidationSpecification

Details about configurations for one or more training jobs that SageMaker runs to test the algorithm.

Type: AlgorithmValidationSpecification object

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeApp

Service: Amazon SageMaker Service

Describes the app.

Request Syntax

```
{
   "AppName": "string",
   "AppType": "string",
   "DomainId": "string",
   "SpaceName": "string",
   "UserProfileName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AppName

The name of the app.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AppType

The type of app.

Type: String

Valid Values: `JupyterServer | KernelGateway`

Required: Yes

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

SpaceName

The name of the space.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

UserProfileName

The user profile name. If this value is not set, then `SpaceName` must be set.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Response Syntax

```
{
   "AppArn": "string",
   "AppName": "string",
   "AppType": "string",
   "BuiltInLifecycleConfigArn": "string",
   "CreationTime": number,
```

```
   "DomainId": "string",
   "EffectiveTrustedIdentityPropagationStatus": "string",
   "FailureReason": "string",
   "LastHealthCheckTimestamp": number,
   "LastUserActivityTimestamp": number,
   "RecoveryMode": boolean,
   "ResourceSpec": {
      "InstanceType": "string",
      "LifecycleConfigArn": "string",
      "SageMakerImageArn": "string",
      "SageMakerImageVersionAlias": "string",
      "SageMakerImageVersionArn": "string",
      "TrainingPlanArn": "string"
   },
   "SpaceName": "string",
   "Status": "string",
   "UserProfileName": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AppArn

The Amazon Resource Name (ARN) of the app.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:app/.*`

AppName

The name of the app.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

AppType

The type of app.

Type: String

Valid Values: `JupyterServer | KernelGateway`

BuiltInLifecycleConfigArn

The lifecycle configuration that runs before the default lifecycle configuration

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:studio-lifecycle-config/.*|None)`

CreationTime

The creation time of the application.

> **Note**
>
> After an application has been shut down for 24 hours, SageMaker AI deletes all metadata for the application. To be considered an update and retain application metadata, applications must be restarted within 24 hours after the previous application has been shut down. After this time window, creation of an application is considered a new application rather than an update of the previous application.

Type: Timestamp

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

EffectiveTrustedIdentityPropagationStatus

The effective status of Trusted Identity Propagation (TIP) for this application. When enabled, user identities from IAM Identity Center are being propagated through the application to TIP enabled AWS services. When disabled, standard IAM role-based access is used.

Type: String

Valid Values: `ENABLED | DISABLED`

FailureReason

The failure reason.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

LastHealthCheckTimestamp

The timestamp of the last health check.

Type: Timestamp

LastUserActivityTimestamp

The timestamp of the last user's activity. `LastUserActivityTimestamp` is also updated when SageMaker AI performs health checks without user activity. As a result, this value is set to the same value as `LastHealthCheckTimestamp`.

Type: Timestamp

RecoveryMode

Indicates whether the application is launched in recovery mode.

Type: Boolean

ResourceSpec

The instance type and the Amazon Resource Name (ARN) of the SageMaker AI image created on the instance.

Type: ResourceSpec object

SpaceName

The name of the space. If this value is not set, then `UserProfileName` must be set.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Status

The status.

Type: String

Valid Values: `Deleted | Deleting | Failed | InService | Pending`

UserProfileName

The user profile name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeAppImageConfig

Service: Amazon SageMaker Service

Describes an AppImageConfig.

Request Syntax

```
{
   "AppImageConfigName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AppImageConfigName

The name of the AppImageConfig to describe.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AppImageConfigArn": "string",
   "AppImageConfigName": "string",
   "CodeEditorAppImageConfig": {
      "ContainerConfig": {
         "ContainerArguments": [ "string" ],
         "ContainerEntrypoint": [ "string" ],
         "ContainerEnvironmentVariables": {
            "string" : "string"
```

```
            }
         },
         "FileSystemConfig": {
            "DefaultGid": number,
            "DefaultUid": number,
            "MountPath": "string"
         }
      },
      "CreationTime": number,
      "JupyterLabAppImageConfig": {
         "ContainerConfig": {
            "ContainerArguments": [ "string" ],
            "ContainerEntrypoint": [ "string" ],
            "ContainerEnvironmentVariables": {
               "string" : "string"
            }
         },
         "FileSystemConfig": {
            "DefaultGid": number,
            "DefaultUid": number,
            "MountPath": "string"
         }
      },
      "KernelGatewayImageConfig": {
         "FileSystemConfig": {
            "DefaultGid": number,
            "DefaultUid": number,
            "MountPath": "string"
         },
         "KernelSpecs": [
            {
               "DisplayName": "string",
               "Name": "string"
            }
         ]
      },
      "LastModifiedTime": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AppImageConfigArn

The ARN of the AppImageConfig.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:app-image-config/.*`

AppImageConfigName

The name of the AppImageConfig.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

CodeEditorAppImageConfig

The configuration of the Code Editor app.

Type: CodeEditorAppImageConfig object

CreationTime

When the AppImageConfig was created.

Type: Timestamp

JupyterLabAppImageConfig

The configuration of the JupyterLab app.

Type: JupyterLabAppImageConfig object

KernelGatewayImageConfig

The configuration of a KernelGateway app.

Type: KernelGatewayImageConfig object

LastModifiedTime

When the AppImageConfig was last modified.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeArtifact

Service: Amazon SageMaker Service

Describes an artifact.

Request Syntax

```
{
   "ArtifactArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ArtifactArn

The Amazon Resource Name (ARN) of the artifact to describe.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:artifact/.*`

Required: Yes

Response Syntax

```
{
   "ArtifactArn": "string",
   "ArtifactName": "string",
   "ArtifactType": "string",
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
```

```
      "UserProfileName": "string"
   },
   "CreationTime": number,
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "LineageGroupArn": "string",
   "MetadataProperties": {
      "CommitId": "string",
      "GeneratedBy": "string",
      "ProjectId": "string",
      "Repository": "string"
   },
   "Properties": {
      "string" : "string"
   },
   "Source": {
      "SourceTypes": [
         {
            "SourceIdType": "string",
            "Value": "string"
         }
      ],
      "SourceUri": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ArtifactArn

The Amazon Resource Name (ARN) of the artifact.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:artifact/.*`

ArtifactName

The name of the artifact.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial|experiment-trial-component|artifact|action|context)\/)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,119})`

ArtifactType

The type of the artifact.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

When the artifact was created.

Type: Timestamp

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

When the artifact was last modified.

Type: Timestamp

LineageGroupArn

The Amazon Resource Name (ARN) of the lineage group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:lineage-group/.*`

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Properties

A list of the artifact's properties.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 30 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2500.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Source

The source of the artifact.

Type: ArtifactSource object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeAutoMLJob

Service: Amazon SageMaker Service

Returns information about an AutoML job created by calling CreateAutoMLJob.

> **Note**
>
> AutoML jobs created by calling CreateAutoMLJobV2 cannot be described by `DescribeAutoMLJob`.

Request Syntax

```
{
   "AutoMLJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AutoMLJobName

Requests information about an AutoML job using its unique name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

Response Syntax

```
{
   "AutoMLJobArn": "string",
   "AutoMLJobArtifacts": {
      "CandidateDefinitionNotebookLocation": "string",
```

```
      "DataExplorationNotebookLocation": "string"
   },
   "AutoMLJobConfig": {
      "CandidateGenerationConfig": {
         "AlgorithmsConfig": [
            {
               "AutoMLAlgorithms": [ "string" ]
            }
         ],
         "FeatureSpecificationS3Uri": "string"
      },
      "CompletionCriteria": {
         "MaxAutoMLJobRuntimeInSeconds": number,
         "MaxCandidates": number,
         "MaxRuntimePerTrainingJobInSeconds": number
      },
      "DataSplitConfig": {
         "ValidationFraction": number
      },
      "Mode": "string",
      "SecurityConfig": {
         "EnableInterContainerTrafficEncryption": boolean,
         "VolumeKmsKeyId": "string",
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      }
   },
   "AutoMLJobName": "string",
   "AutoMLJobObjective": {
      "MetricName": "string"
   },
   "AutoMLJobSecondaryStatus": "string",
   "AutoMLJobStatus": "string",
   "BestCandidate": {
      "CandidateName": "string",
      "CandidateProperties": {
         "CandidateArtifactLocations": {
            "BacktestResults": "string",
            "Explainability": "string",
            "ModelInsights": "string"
         },
         "CandidateMetrics": [
```

```
            {
               "MetricName": "string",
               "Set": "string",
               "StandardMetricName": "string",
               "Value": number
            }
         ]
      },
      "CandidateStatus": "string",
      "CandidateSteps": [
         {
            "CandidateStepArn": "string",
            "CandidateStepName": "string",
            "CandidateStepType": "string"
         }
      ],
      "CreationTime": number,
      "EndTime": number,
      "FailureReason": "string",
      "FinalAutoMLJobObjectiveMetric": {
         "MetricName": "string",
         "StandardMetricName": "string",
         "Type": "string",
         "Value": number
      },
      "InferenceContainerDefinitions": {
         "string" : [
            {
               "Environment": {
                  "string" : "string"
               },
               "Image": "string",
               "ModelDataUrl": "string"
            }
         ]
      },
      "InferenceContainers": [
         {
            "Environment": {
               "string" : "string"
            },
            "Image": "string",
            "ModelDataUrl": "string"
         }
```

```
      ],
      "LastModifiedTime": number,
      "ObjectiveStatus": "string"
   },
   "CreationTime": number,
   "EndTime": number,
   "FailureReason": "string",
   "GenerateCandidateDefinitionsOnly": boolean,
   "InputDataConfig": [
      {
         "ChannelType": "string",
         "CompressionType": "string",
         "ContentType": "string",
         "DataSource": {
            "S3DataSource": {
               "S3DataType": "string",
               "S3Uri": "string"
            }
         },
         "SampleWeightAttributeName": "string",
         "TargetAttributeName": "string"
      }
   ],
   "LastModifiedTime": number,
   "ModelDeployConfig": {
      "AutoGenerateEndpointName": boolean,
      "EndpointName": "string"
   },
   "ModelDeployResult": {
      "EndpointName": "string"
   },
   "OutputDataConfig": {
      "KmsKeyId": "string",
      "S3OutputPath": "string"
   },
   "PartialFailureReasons": [
      {
         "PartialFailureMessage": "string"
      }
   ],
   "ProblemType": "string",
   "ResolvedAttributes": {
      "AutoMLJobObjective": {
         "MetricName": "string"
```

```
      },
      "CompletionCriteria": {
         "MaxAutoMLJobRuntimeInSeconds": number,
         "MaxCandidates": number,
         "MaxRuntimePerTrainingJobInSeconds": number
      },
      "ProblemType": "string"
   },
   "RoleArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AutoMLJobArn

Returns the ARN of the AutoML job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

AutoMLJobArtifacts

Returns information on the job's artifacts found in `AutoMLJobArtifacts`.

Type: AutoMLJobArtifacts object

AutoMLJobConfig

Returns the configuration for the AutoML job.

Type: AutoMLJobConfig object

AutoMLJobName

Returns the name of the AutoML job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

AutoMLJobObjective

Returns the job's objective.

Type: AutoMLJobObjective object

AutoMLJobSecondaryStatus

Returns the secondary status of the AutoML job.

Type: String

Valid Values: `Starting | MaxCandidatesReached | Failed | Stopped | MaxAutoMLJobRuntimeReached | Stopping | CandidateDefinitionsGenerated | Completed | ExplainabilityError | DeployingModel | ModelDeploymentError | GeneratingModelInsightsReport | ModelInsightsError | AnalyzingData | FeatureEngineering | ModelTuning | GeneratingExplainabilityReport | TrainingModels | PreTraining`

AutoMLJobStatus

Returns the status of the AutoML job.

Type: String

Valid Values: `Completed | InProgress | Failed | Stopped | Stopping`

BestCandidate

The best model candidate selected by SageMaker AI Autopilot using both the best objective metric and lowest InferenceLatency for an experiment.

Type: AutoMLCandidate object

CreationTime

Returns the creation time of the AutoML job.

Type: Timestamp

EndTime

Returns the end time of the AutoML job.

Type: Timestamp

FailureReason

Returns the failure reason for an AutoML job, when applicable.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

GenerateCandidateDefinitionsOnly

Indicates whether the output for an AutoML job generates candidate definitions only.

Type: Boolean

InputDataConfig

Returns the input data configuration for the AutoML job.

Type: Array of AutoMLChannel objects

Array Members: Minimum number of 1 item. Maximum number of 2 items.

LastModifiedTime

Returns the job's last modified time.

Type: Timestamp

ModelDeployConfig

Indicates whether the model was deployed automatically to an endpoint and the name of that endpoint if deployed automatically.

Type: ModelDeployConfig object

ModelDeployResult

Provides information about endpoint for the model deployment.

Type: ModelDeployResult object

OutputDataConfig

Returns the job's output data config.

Type: AutoMLOutputDataConfig object

PartialFailureReasons

Returns a list of reasons for partial failures within an AutoML job.

Type: Array of AutoMLPartialFailureReason objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

ProblemType

Returns the job's problem type.

Type: String

Valid Values: `BinaryClassification | MulticlassClassification | Regression`

ResolvedAttributes

Contains `ProblemType`, `AutoMLJobObjective`, and `CompletionCriteria`. If you do not provide these values, they are inferred.

Type: ResolvedAttributes object

RoleArn

The ARN of the IAM role that has read permission to the input data location and write permission to the output data location in Amazon S3.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeAutoMLJobV2

Service: Amazon SageMaker Service

Returns information about an AutoML job created by calling CreateAutoMLJobV2 or CreateAutoMLJob.

Request Syntax

```
{
   "AutoMLJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AutoMLJobName

Requests information about an AutoML job V2 using its unique name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

Response Syntax

```
{
   "AutoMLComputeConfig": {
      "EmrServerlessComputeConfig": {
         "ExecutionRoleARN": "string"
      }
   },
   "AutoMLJobArn": "string",
   "AutoMLJobArtifacts": {
      "CandidateDefinitionNotebookLocation": "string",
      "DataExplorationNotebookLocation": "string"
```

```
   },
   "AutoMLJobInputDataConfig": [
      {
         "ChannelType": "string",
         "CompressionType": "string",
         "ContentType": "string",
         "DataSource": {
            "S3DataSource": {
               "S3DataType": "string",
               "S3Uri": "string"
            }
         }
      }
   ],
   "AutoMLJobName": "string",
   "AutoMLJobObjective": {
      "MetricName": "string"
   },
   "AutoMLJobSecondaryStatus": "string",
   "AutoMLJobStatus": "string",
   "AutoMLProblemTypeConfig": { ... },
   "AutoMLProblemTypeConfigName": "string",
   "BestCandidate": {
      "CandidateName": "string",
      "CandidateProperties": {
         "CandidateArtifactLocations": {
            "BacktestResults": "string",
            "Explainability": "string",
            "ModelInsights": "string"
         },
         "CandidateMetrics": [
            {
               "MetricName": "string",
               "Set": "string",
               "StandardMetricName": "string",
               "Value": number
            }
         ]
      },
      "CandidateStatus": "string",
      "CandidateSteps": [
         {
            "CandidateStepArn": "string",
            "CandidateStepName": "string",
```

```
            "CandidateStepType": "string"
         }
      ],
      "CreationTime": number,
      "EndTime": number,
      "FailureReason": "string",
      "FinalAutoMLJobObjectiveMetric": {
         "MetricName": "string",
         "StandardMetricName": "string",
         "Type": "string",
         "Value": number
      },
      "InferenceContainerDefinitions": {
         "string" : [
            {
               "Environment": {
                  "string" : "string"
               },
               "Image": "string",
               "ModelDataUrl": "string"
            }
         ]
      },
      "InferenceContainers": [
         {
            "Environment": {
               "string" : "string"
            },
            "Image": "string",
            "ModelDataUrl": "string"
         }
      ],
      "LastModifiedTime": number,
      "ObjectiveStatus": "string"
   },
   "CreationTime": number,
   "DataSplitConfig": {
      "ValidationFraction": number
   },
   "EndTime": number,
   "FailureReason": "string",
   "LastModifiedTime": number,
   "ModelDeployConfig": {
      "AutoGenerateEndpointName": boolean,
```

```
      "EndpointName": "string"
   },
   "ModelDeployResult": {
      "EndpointName": "string"
   },
   "OutputDataConfig": {
      "KmsKeyId": "string",
      "S3OutputPath": "string"
   },
   "PartialFailureReasons": [
      {
         "PartialFailureMessage": "string"
      }
   ],
   "ResolvedAttributes": {
      "AutoMLJobObjective": {
         "MetricName": "string"
      },
      "AutoMLProblemTypeResolvedAttributes": { ... },
      "CompletionCriteria": {
         "MaxAutoMLJobRuntimeInSeconds": number,
         "MaxCandidates": number,
         "MaxRuntimePerTrainingJobInSeconds": number
      }
   },
   "RoleArn": "string",
   "SecurityConfig": {
      "EnableInterContainerTrafficEncryption": boolean,
      "VolumeKmsKeyId": "string",
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AutoMLComputeConfig

The compute configuration used for the AutoML job V2.

Type: AutoMLComputeConfig object

AutoMLJobArn

Returns the Amazon Resource Name (ARN) of the AutoML job V2.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

AutoMLJobArtifacts

The artifacts that are generated during an AutoML job.

Type: AutoMLJobArtifacts object

AutoMLJobInputDataConfig

Returns an array of channel objects describing the input data and their location.

Type: Array of AutoMLJobChannel objects

Array Members: Minimum number of 1 item. Maximum number of 2 items.

AutoMLJobName

Returns the name of the AutoML job V2.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

AutoMLJobObjective

Returns the job's objective.

Type: AutoMLJobObjective object

AutoMLJobSecondaryStatus

Returns the secondary status of the AutoML job V2.

Type: String

Valid Values: `Starting | MaxCandidatesReached | Failed | Stopped | MaxAutoMLJobRuntimeReached | Stopping | CandidateDefinitionsGenerated | Completed | ExplainabilityError | DeployingModel | ModelDeploymentError | GeneratingModelInsightsReport | ModelInsightsError | AnalyzingData | FeatureEngineering | ModelTuning | GeneratingExplainabilityReport | TrainingModels | PreTraining`

AutoMLJobStatus

Returns the status of the AutoML job V2.

Type: String

Valid Values: `Completed | InProgress | Failed | Stopped | Stopping`

AutoMLProblemTypeConfig

Returns the configuration settings of the problem type set for the AutoML job V2.

Type: AutoMLProblemTypeConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

AutoMLProblemTypeConfigName

Returns the name of the problem type configuration set for the AutoML job V2.

Type: String

Valid Values: `ImageClassification | TextClassification | TimeSeriesForecasting | Tabular | TextGeneration`

BestCandidate

Information about the candidate produced by an AutoML training job V2, including its status, steps, and other properties.

Type: AutoMLCandidate object

CreationTime

Returns the creation time of the AutoML job V2.

Type: Timestamp

DataSplitConfig

Returns the configuration settings of how the data are split into train and validation datasets.

Type: AutoMLDataSplitConfig object

EndTime

Returns the end time of the AutoML job V2.

Type: Timestamp

FailureReason

Returns the reason for the failure of the AutoML job V2, when applicable.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

LastModifiedTime

Returns the job's last modified time.

Type: Timestamp

ModelDeployConfig

Indicates whether the model was deployed automatically to an endpoint and the name of that endpoint if deployed automatically.

Type: ModelDeployConfig object

ModelDeployResult

Provides information about endpoint for the model deployment.

Type: ModelDeployResult object

OutputDataConfig

Returns the job's output data config.

Type: AutoMLOutputDataConfig object

PartialFailureReasons

Returns a list of reasons for partial failures within an AutoML job V2.

Type: Array of AutoMLPartialFailureReason objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

ResolvedAttributes

Returns the resolved attributes used by the AutoML job V2.

Type: AutoMLResolvedAttributes object

RoleArn

The ARN of the IAM role that has read permission to the input data location and write permission to the output data location in Amazon S3.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

SecurityConfig

Returns the security configuration for traffic encryption or Amazon VPC settings.

Type: AutoMLSecurityConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeCluster

Service: Amazon SageMaker Service

Retrieves information of a SageMaker HyperPod cluster.

Request Syntax

```
{
   "ClusterName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterName

The string name or the Amazon Resource Name (ARN) of the SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

Response Syntax

```
{
   "AutoScaling": {
      "AutoScalerType": "string",
      "FailureMessage": "string",
      "Mode": "string",
      "Status": "string"
   },
   "ClusterArn": "string",
   "ClusterName": "string",
   "ClusterRole": "string",
```

```
   "ClusterStatus": "string",
   "CreationTime": number,
   "FailureMessage": "string",
   "InstanceGroups": [
      {
         "ActiveOperations": {
            "string" : number
         },
         "ActiveSoftwareUpdateConfig": {
            "AutoRollbackConfiguration": [
               {
                  "AlarmName": "string"
               }
            ],
            "RollingUpdatePolicy": {
               "MaximumBatchSize": {
                  "Type": "string",
                  "Value": number
               },
               "RollbackMaximumBatchSize": {
                  "Type": "string",
                  "Value": number
               }
            },
            "WaitIntervalInSeconds": number
         },
         "CapacityRequirements": {
            "OnDemand": {
            },
            "Spot": {
            }
         },
         "CurrentCount": number,
         "CurrentImageId": "string",
         "DesiredImageId": "string",
         "ExecutionRole": "string",
         "ImageVersionStatus": "string",
         "InstanceGroupName": "string",
         "InstanceRequirements": {
            "CurrentInstanceTypes": [ "string" ],
            "DesiredInstanceTypes": [ "string" ]
         },
         "InstanceStorageConfigs": [
            { ... }
```

```
        ],
        "InstanceType": "string",
        "InstanceTypeDetails": [
           {
              "CurrentCount": number,
              "InstanceType": "string",
              "ThreadsPerCore": number
           }
        ],
        "KubernetesConfig": {
           "CurrentLabels": {
              "string" : "string"
           },
           "CurrentTaints": [
              {
                 "Effect": "string",
                 "Key": "string",
                 "Value": "string"
              }
           ],
           "DesiredLabels": {
              "string" : "string"
           },
           "DesiredTaints": [
              {
                 "Effect": "string",
                 "Key": "string",
                 "Value": "string"
              }
           ]
        },
        "LifeCycleConfig": {
           "OnCreate": "string",
           "OnInitComplete": "string",
           "SourceS3Uri": "string"
        },
        "MinCount": number,
        "NetworkInterface": {
           "InterfaceType": "string"
        },
        "OnStartDeepHealthChecks": [ "string" ],
        "OverrideVpcConfig": {
           "SecurityGroupIds": [ "string" ],
           "Subnets": [ "string" ]
```

```
         },
         "ScheduledUpdateConfig": {
            "DeploymentConfig": {
               "AutoRollbackConfiguration": [
                  {
                     "AlarmName": "string"
                  }
               ],
               "RollingUpdatePolicy": {
                  "MaximumBatchSize": {
                     "Type": "string",
                     "Value": number
                  },
                  "RollbackMaximumBatchSize": {
                     "Type": "string",
                     "Value": number
                  }
               },
               "WaitIntervalInSeconds": number
            },
            "ScheduleExpression": "string"
         },
         "SlurmConfig": {
            "NodeType": "string",
            "PartitionNames": [ "string" ]
         },
         "SoftwareUpdateStatus": "string",
         "Status": "string",
         "TargetCount": number,
         "TargetStateCount": number,
         "ThreadsPerCore": number,
         "TrainingPlanArn": "string",
         "TrainingPlanStatus": "string"
      }
   ],
   "NodeProvisioningMode": "string",
   "NodeRecovery": "string",
   "Orchestrator": {
      "Eks": {
         "ClusterArn": "string"
      },
      "Slurm": {
         "SlurmConfigStrategy": "string"
      }
```

```
   },
   "RestrictedInstanceGroups": [
      {
         "CurrentCount": number,
         "EnvironmentConfig": {
            "FSxLustreConfig": {
               "PerUnitStorageThroughput": number,
               "SizeInGiB": number
            },
            "S3OutputPath": "string"
         },
         "ExecutionRole": "string",
         "InstanceGroupName": "string",
         "InstanceStorageConfigs": [
            { ... }
         ],
         "InstanceType": "string",
         "OnStartDeepHealthChecks": [ "string" ],
         "OverrideVpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         },
         "ScheduledUpdateConfig": {
            "DeploymentConfig": {
               "AutoRollbackConfiguration": [
                  {
                     "AlarmName": "string"
                  }
               ],
               "RollingUpdatePolicy": {
                  "MaximumBatchSize": {
                     "Type": "string",
                     "Value": number
                  },
                  "RollbackMaximumBatchSize": {
                     "Type": "string",
                     "Value": number
                  }
               },
               "WaitIntervalInSeconds": number
            },
            "ScheduleExpression": "string"
         },
         "Status": "string",
```

```
            "TargetCount": number,
            "ThreadsPerCore": number,
            "TrainingPlanArn": "string",
            "TrainingPlanStatus": "string"
         }
      ],
      "RestrictedInstanceGroupsConfig": {
         "SharedEnvironmentConfig": {
            "CurrentFSxLustreConfig": {
               "PerUnitStorageThroughput": number,
               "SizeInGiB": number
            },
            "CurrentFSxLustreDeletionPolicy": "string",
            "DesiredFSxLustreConfig": {
               "PerUnitStorageThroughput": number,
               "SizeInGiB": number
            },
            "DesiredFSxLustreDeletionPolicy": "string"
         }
      },
      "TieredStorageConfig": {
         "InstanceMemoryAllocationPercentage": number,
         "Mode": "string"
      },
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AutoScaling

The current autoscaling configuration and status for the autoscaler.

Type: ClusterAutoScalingConfigOutput object

ClusterArn

The Amazon Resource Name (ARN) of the SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

ClusterName

The name of the SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

ClusterRole

The Amazon Resource Name (ARN) of the IAM role that HyperPod uses for cluster autoscaling operations.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

ClusterStatus

The status of the SageMaker HyperPod cluster.

Type: String

Valid Values: `Creating | Deleting | Failed | InService | RollingBack | SystemUpdating | Updating`

CreationTime

The time when the SageMaker Cluster is created.

Type: Timestamp

FailureMessage

The failure message of the SageMaker HyperPod cluster.

Type: String

InstanceGroups

The instance groups of the SageMaker HyperPod cluster.

Type: Array of ClusterInstanceGroupDetails objects

NodeProvisioningMode

The mode used for provisioning nodes in the cluster.

Type: String

Valid Values: `Continuous`

NodeRecovery

The node recovery mode configured for the SageMaker HyperPod cluster.

Type: String

Valid Values: `Automatic | None`

Orchestrator

The type of orchestrator used for the SageMaker HyperPod cluster.

Type: ClusterOrchestrator object

RestrictedInstanceGroups

The specialized instance groups for training models like Amazon Nova to be created in the SageMaker HyperPod cluster.

Type: Array of ClusterRestrictedInstanceGroupDetails objects

RestrictedInstanceGroupsConfig

The configuration for the restricted instance groups (RIG) in the SageMaker HyperPod cluster.

Type: ClusterRestrictedInstanceGroupsConfigOutput object

TieredStorageConfig

The current configuration for managed tier checkpointing on the HyperPod cluster. For example, this shows whether the feature is enabled and the percentage of cluster memory allocated for checkpoint storage.

Type: ClusterTieredStorageConfig object

VpcConfig

Specifies an Amazon Virtual Private Cloud (VPC) that your SageMaker jobs, hosted models, and compute resources have access to. You can control access to and from your resources by configuring a VPC. For more information, see Give SageMaker Access to Resources in your Amazon VPC.

Type: VpcConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeClusterEvent

Service: Amazon SageMaker Service

Retrieves detailed information about a specific event for a given HyperPod cluster. This functionality is only supported when the `NodeProvisioningMode` is set to `Continuous`.

Request Syntax

```
{
   "ClusterName": "string",
   "EventId": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterName

The name or Amazon Resource Name (ARN) of the HyperPod cluster associated with the event.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

EventId

The unique identifier (UUID) of the event to describe. This ID can be obtained from the `ListClusterEvents` operation.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

Response Syntax

```
{
   "EventDetails": {
      "ClusterArn": "string",
      "ClusterName": "string",
      "Description": "string",
      "EventDetails": {
         "EventMetadata": { ... }
      },
      "EventId": "string",
      "EventLevel": "string",
      "EventTime": number,
      "InstanceGroupName": "string",
      "InstanceId": "string",
      "ResourceType": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

EventDetails

Detailed information about the requested cluster event, including event metadata for various resource types such as `Cluster`, `InstanceGroup`, `Instance`, and their associated attributes.

Type: ClusterEventDetail object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeClusterNode

Service: Amazon SageMaker Service

Retrieves information of a node (also called a *instance* interchangeably) of a SageMaker HyperPod cluster.

Request Syntax

```
{
   "ClusterName": "string",
   "NodeId": "string",
   "NodeLogicalId": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterName

The string name or the Amazon Resource Name (ARN) of the SageMaker HyperPod cluster in which the node is.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

NodeId

The ID of the SageMaker HyperPod cluster node.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: No

NodeLogicalId

The logical identifier of the node to describe. You can specify either `NodeLogicalId` or `InstanceId`, but not both. `NodeLogicalId` can be used to describe nodes that are still being provisioned and don't yet have an `InstanceId` assigned.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Required: No

Response Syntax

```
{
   "NodeDetails": {
      "CapacityType": "string",
      "CurrentImageId": "string",
      "DesiredImageId": "string",
      "ImageVersionStatus": "string",
      "InstanceGroupName": "string",
      "InstanceId": "string",
      "InstanceStatus": {
         "Message": "string",
         "Status": "string"
      },
      "InstanceStorageConfigs": [
         { ... }
      ],
      "InstanceType": "string",
      "KubernetesConfig": {
         "CurrentLabels": {
            "string" : "string"
         },
         "CurrentTaints": [
            {
               "Effect": "string",
               "Key": "string",
               "Value": "string"
```

```
            }
         ],
         "DesiredLabels": {
            "string" : "string"
         },
         "DesiredTaints": [
            {
               "Effect": "string",
               "Key": "string",
               "Value": "string"
            }
         ]
      },
      "LastSoftwareUpdateTime": number,
      "LaunchTime": number,
      "LifeCycleConfig": {
         "OnCreate": "string",
         "OnInitComplete": "string",
         "SourceS3Uri": "string"
      },
      "NetworkInterface": {
         "InterfaceType": "string"
      },
      "NodeLogicalId": "string",
      "OverrideVpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      },
      "Placement": {
         "AvailabilityZone": "string",
         "AvailabilityZoneId": "string"
      },
      "PrivateDnsHostname": "string",
      "PrivatePrimaryIp": "string",
      "PrivatePrimaryIpv6": "string",
      "ThreadsPerCore": number,
      "UltraServerInfo": {
         "Id": "string",
         "Type": "string"
      }
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NodeDetails

The details of the SageMaker HyperPod cluster node.

Type: ClusterNodeDetails object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeClusterSchedulerConfig

Service: Amazon SageMaker Service

Description of the cluster policy. This policy is used for task prioritization and fair-share allocation. This helps prioritize critical workloads and distributes idle compute across entities.

Request Syntax

```
{
   "ClusterSchedulerConfigId": "string",
   "ClusterSchedulerConfigVersion": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterSchedulerConfigId

ID of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[a-z0-9]{12}`

Required: Yes

ClusterSchedulerConfigVersion

Version of the cluster policy.

Type: Integer

Required: No

Response Syntax

```
{
   "ClusterArn": "string",
```

```
   "ClusterSchedulerConfigArn": "string",
   "ClusterSchedulerConfigId": "string",
   "ClusterSchedulerConfigVersion": number,
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "CreationTime": number,
   "Description": "string",
   "FailureReason": "string",
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "Name": "string",
   "SchedulerConfig": {
      "FairShare": "string",
      "IdleResourceSharing": "string",
      "PriorityClasses": [
         {
            "Name": "string",
            "Weight": number
         }
      ]
   },
   "Status": "string",
   "StatusDetails": {
      "string" : "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ClusterArn

ARN of the cluster where the cluster policy is applied.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

ClusterSchedulerConfigArn

ARN of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:cluster-scheduler-config/[a-z0-9]{12}`

ClusterSchedulerConfigId

ID of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[a-z0-9]{12}`

ClusterSchedulerConfigVersion

Version of the cluster policy.

Type: Integer

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

Creation time of the cluster policy.

Type: Timestamp

Description

Description of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

FailureReason

Failure reason of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

Last modified time of the cluster policy.

Type: Timestamp

Name

Name of the cluster policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

SchedulerConfig

Cluster policy configuration. This policy is used for task prioritization and fair-share allocation. This helps prioritize critical workloads and distributes idle compute across entities.

Type: SchedulerConfig object

Status

Status of the cluster policy.

Type: String

Valid Values: `Creating | CreateFailed | CreateRollbackFailed | Created | Updating | UpdateFailed | UpdateRollbackFailed | Updated | Deleting | DeleteFailed | DeleteRollbackFailed | Deleted`

StatusDetails

Additional details about the status of the cluster policy. This field provides context when the policy is in a non-active state, such as during creation, updates, or if failures occur.

Type: String to string map

Valid Keys: `PriorityClasses | FairShare | IdleResourceSharing`

Valid Values: `Creating | CreateFailed | CreateRollbackFailed | Created | Updating | UpdateFailed | UpdateRollbackFailed | Updated | Deleting | DeleteFailed | DeleteRollbackFailed | Deleted`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeCodeRepository

Service: Amazon SageMaker Service

Gets details about the specified Git repository.

Request Syntax

```
{
   "CodeRepositoryName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CodeRepositoryName

The name of the Git repository to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CodeRepositoryArn": "string",
   "CodeRepositoryName": "string",
   "CreationTime": number,
   "GitConfig": {
      "Branch": "string",
      "RepositoryUrl": "string",
      "SecretArn": "string"
   },
   "LastModifiedTime": number
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CodeRepositoryArn

The Amazon Resource Name (ARN) of the Git repository.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:code-repository/[\S]{1,2048}`

CodeRepositoryName

The name of the Git repository.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

CreationTime

The date and time that the repository was created.

Type: Timestamp

GitConfig

Configuration details about the repository, including the URL where the repository is located, the default branch, and the Amazon Resource Name (ARN) of the AWS Secrets Manager secret that contains the credentials used to access the repository.

Type: GitConfig object

LastModifiedTime

The date and time that the repository was last changed.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeCompilationJob

Service: Amazon SageMaker Service

Returns information about a model compilation job.

To create a model compilation job, use CreateCompilationJob. To get information about multiple model compilation jobs, use ListCompilationJobs.

Request Syntax

```
{
   "CompilationJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CompilationJobName

The name of the model compilation job that you want information about.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CompilationEndTime": number,
   "CompilationJobArn": "string",
   "CompilationJobName": "string",
   "CompilationJobStatus": "string",
   "CompilationStartTime": number,
   "CreationTime": number,
   "DerivedInformation": {
```

```
      "DerivedDataInputConfig": "string"
   },
   "FailureReason": "string",
   "InferenceImage": "string",
   "InputConfig": {
      "DataInputConfig": "string",
      "Framework": "string",
      "FrameworkVersion": "string",
      "S3Uri": "string"
   },
   "LastModifiedTime": number,
   "ModelArtifacts": {
      "S3ModelArtifacts": "string"
   },
   "ModelDigests": {
      "ArtifactDigest": "string"
   },
   "ModelPackageVersionArn": "string",
   "OutputConfig": {
      "CompilerOptions": "string",
      "KmsKeyId": "string",
      "S3OutputLocation": "string",
      "TargetDevice": "string",
      "TargetPlatform": {
         "Accelerator": "string",
         "Arch": "string",
         "Os": "string"
      }
   },
   "RoleArn": "string",
   "StoppingCondition": {
      "MaxPendingTimeInSeconds": number,
      "MaxRuntimeInSeconds": number,
      "MaxWaitTimeInSeconds": number
   },
   "VpcConfig": {
      "SecurityGroupIds": [ "string" ],
      "Subnets": [ "string" ]
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CompilationEndTime

The time when the model compilation job on a compilation job instance ended. For a successful or stopped job, this is when the job's model artifacts have finished uploading. For a failed job, this is when Amazon SageMaker AI detected that the job failed.

Type: Timestamp

CompilationJobArn

The Amazon Resource Name (ARN) of the model compilation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:compilation-job/.*`

CompilationJobName

The name of the model compilation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

CompilationJobStatus

The status of the model compilation job.

Type: String

Valid Values: `INPROGRESS | COMPLETED | FAILED | STARTING | STOPPING | STOPPED`

CompilationStartTime

The time when the model compilation job started the `CompilationJob` instances.

You are billed for the time between this timestamp and the timestamp in the `CompilationEndTime` field. In Amazon CloudWatch Logs, the start time might be later than this time. That's because it takes time to download the compilation job, which depends on the size of the compilation job container.

Type: Timestamp

CreationTime

The time that the model compilation job was created.

Type: Timestamp

DerivedInformation

Information that SageMaker Neo automatically derived about the model.

Type: DerivedInformation object

FailureReason

If a model compilation job failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

InferenceImage

The inference image to use when compiling a model. Specify an image only if the target device is a cloud instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

InputConfig

Information about the location in Amazon S3 of the input model artifacts, the name and shape of the expected data inputs, and the framework in which the model was trained.

Type: InputConfig object

LastModifiedTime

The time that the status of the model compilation job was last modified.

Type: Timestamp

ModelArtifacts

Information about the location in Amazon S3 that has been configured for storing the model artifacts used in the compilation job.

Type: ModelArtifacts object

ModelDigests

Provides a BLAKE2 hash value that identifies the compiled model artifacts in Amazon S3.

Type: ModelDigests object

ModelPackageVersionArn

The Amazon Resource Name (ARN) of the versioned model package that was provided to SageMaker Neo when you initiated a compilation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

OutputConfig

Information about the output location for the compiled model and the target device that the model runs on.

Type: OutputConfig object

RoleArn

The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker AI assumes to perform the model compilation job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

StoppingCondition

Specifies a limit to how long a model compilation job can run. When the job reaches the time limit, Amazon SageMaker AI ends the compilation job. Use this API to cap model training costs.

Type: StoppingCondition object

VpcConfig

A VpcConfig object that specifies the VPC that you want your compilation job to connect to. Control access to your models by configuring the VPC. For more information, see Protect Compilation Jobs by Using an Amazon Virtual Private Cloud.

Type: NeoVpcConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeComputeQuota

Service: Amazon SageMaker Service

Description of the compute allocation definition.

Request Syntax

```
{
   "ComputeQuotaId": "string",
   "ComputeQuotaVersion": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ComputeQuotaId

ID of the compute allocation definition.

Type: String

Pattern: `[a-z0-9]{12}`

Required: Yes

ComputeQuotaVersion

Version of the compute allocation definition.

Type: Integer

Required: No

Response Syntax

```
{
   "ActivationState": "string",
   "ClusterArn": "string",
   "ComputeQuotaArn": "string",
   "ComputeQuotaConfig": {
```

```
      "ComputeQuotaResources": [
         {
            "AcceleratorPartition": {
               "Count": number,
               "Type": "string"
            },
            "Accelerators": number,
            "Count": number,
            "InstanceType": "string",
            "MemoryInGiB": number,
            "VCpu": number
         }
      ],
      "PreemptTeamTasks": "string",
      "ResourceSharingConfig": {
         "AbsoluteBorrowLimits": [
            {
               "AcceleratorPartition": {
                  "Count": number,
                  "Type": "string"
               },
               "Accelerators": number,
               "Count": number,
               "InstanceType": "string",
               "MemoryInGiB": number,
               "VCpu": number
            }
         ],
         "BorrowLimit": number,
         "Strategy": "string"
      }
   },
   "ComputeQuotaId": "string",
   "ComputeQuotaTarget": {
      "FairShareWeight": number,
      "TeamName": "string"
   },
   "ComputeQuotaVersion": number,
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
```

```
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "CreationTime": number,
   "Description": "string",
   "FailureReason": "string",
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "Name": "string",
   "Status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ActivationState

The state of the compute allocation being described. Use to enable or disable compute allocation.

Default is `Enabled`.

Type: String

Valid Values: `Enabled | Disabled`

ClusterArn

ARN of the cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

ComputeQuotaArn

ARN of the compute allocation definition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:compute-quota/[a-z0-9]{12}`

ComputeQuotaConfig

Configuration of the compute allocation definition. This includes the resource sharing option, and the setting to preempt low priority tasks.

Type: ComputeQuotaConfig object

ComputeQuotaId

ID of the compute allocation definition.

Type: String

Pattern: `[a-z0-9]{12}`

ComputeQuotaTarget

The target entity to allocate compute resources to.

Type: ComputeQuotaTarget object

ComputeQuotaVersion

Version of the compute allocation definition.

Type: Integer

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

Creation time of the compute allocation configuration.

Type: Timestamp

Description

Description of the compute allocation definition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

FailureReason

Failure reason of the compute allocation definition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

Last modified time of the compute allocation configuration.

Type: Timestamp

Name

Name of the compute allocation definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Status

Status of the compute allocation definition.

Type: String

Valid Values: `Creating | CreateFailed | CreateRollbackFailed | Created | Updating | UpdateFailed | UpdateRollbackFailed | Updated | Deleting | DeleteFailed | DeleteRollbackFailed | Deleted`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeContext

Service: Amazon SageMaker Service

Describes a context.

Request Syntax

```
{
   "ContextName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ContextName

The name of the context to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:context\/)?([a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,119})`

Required: Yes

Response Syntax

```
{
   "ContextArn": "string",
   "ContextName": "string",
   "ContextType": "string",
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
```

```
            "SourceIdentity": "string"
         },
         "UserProfileArn": "string",
         "UserProfileName": "string"
      },
      "CreationTime": number,
      "Description": "string",
      "LastModifiedBy": {
         "DomainId": "string",
         "IamIdentity": {
            "Arn": "string",
            "PrincipalId": "string",
            "SourceIdentity": "string"
         },
         "UserProfileArn": "string",
         "UserProfileName": "string"
      },
      "LastModifiedTime": number,
      "LineageGroupArn": "string",
      "Properties": {
         "string" : "string"
      },
      "Source": {
         "SourceId": "string",
         "SourceType": "string",
         "SourceUri": "string"
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ContextArn

The Amazon Resource Name (ARN) of the context.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:context/.*`

ContextName

The name of the context.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,119}`

ContextType

The type of the context.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

When the context was created.

Type: Timestamp

Description

The description of the context.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

When the context was last modified.

Type: Timestamp

LineageGroupArn

The Amazon Resource Name (ARN) of the lineage group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:lineage-group/.*`

Properties

A list of the context's properties.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 30 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2500.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Source

The source of the context.

Type: ContextSource object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeDataQualityJobDefinition

Service: Amazon SageMaker Service

Gets the details of a data quality monitoring job definition.

Request Syntax

```
{
   "JobDefinitionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobDefinitionName

The name of the data quality monitoring job definition to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "DataQualityAppSpecification": {
      "ContainerArguments": [ "string" ],
      "ContainerEntrypoint": [ "string" ],
      "Environment": {
         "string" : "string"
      },
      "ImageUri": "string",
      "PostAnalyticsProcessorSourceUri": "string",
      "RecordPreprocessorSourceUri": "string"
   },
```

```
   "DataQualityBaselineConfig": {
      "BaseliningJobName": "string",
      "ConstraintsResource": {
         "S3Uri": "string"
      },
      "StatisticsResource": {
         "S3Uri": "string"
      }
   },
   "DataQualityJobInput": {
      "BatchTransformInput": {
         "DataCapturedDestinationS3Uri": "string",
         "DatasetFormat": {
            "Csv": {
               "Header": boolean
            },
            "Json": {
               "Line": boolean
            },
            "Parquet": {
            }
         },
         "EndTimeOffset": "string",
         "ExcludeFeaturesAttribute": "string",
         "FeaturesAttribute": "string",
         "InferenceAttribute": "string",
         "LocalPath": "string",
         "ProbabilityAttribute": "string",
         "ProbabilityThresholdAttribute": number,
         "S3DataDistributionType": "string",
         "S3InputMode": "string",
         "StartTimeOffset": "string"
      },
      "EndpointInput": {
         "EndpointName": "string",
         "EndTimeOffset": "string",
         "ExcludeFeaturesAttribute": "string",
         "FeaturesAttribute": "string",
         "InferenceAttribute": "string",
         "LocalPath": "string",
         "ProbabilityAttribute": "string",
         "ProbabilityThresholdAttribute": number,
         "S3DataDistributionType": "string",
         "S3InputMode": "string",
```

```
         "StartTimeOffset": "string"
      }
   },
   "DataQualityJobOutputConfig": {
      "KmsKeyId": "string",
      "MonitoringOutputs": [
         {
            "S3Output": {
               "LocalPath": "string",
               "S3UploadMode": "string",
               "S3Uri": "string"
            }
         }
      ]
   },
   "JobDefinitionArn": "string",
   "JobDefinitionName": "string",
   "JobResources": {
      "ClusterConfig": {
         "InstanceCount": number,
         "InstanceType": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
      }
   },
   "NetworkConfig": {
      "EnableInterContainerTrafficEncryption": boolean,
      "EnableNetworkIsolation": boolean,
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
   },
   "RoleArn": "string",
   "StoppingCondition": {
      "MaxRuntimeInSeconds": number
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The time that the data quality monitoring job definition was created.

Type: Timestamp

DataQualityAppSpecification

Information about the container that runs the data quality monitoring job.

Type: DataQualityAppSpecification object

DataQualityBaselineConfig

The constraints and baselines for the data quality monitoring job definition.

Type: DataQualityBaselineConfig object

DataQualityJobInput

The list of inputs for the data quality monitoring job. Currently endpoints are supported.

Type: DataQualityJobInput object

DataQualityJobOutputConfig

The output configuration for monitoring jobs.

Type: MonitoringOutputConfig object

JobDefinitionArn

The Amazon Resource Name (ARN) of the data quality monitoring job definition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

JobDefinitionName

The name of the data quality monitoring job definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

JobResources

Identifies the resources to deploy for a monitoring job.

Type: MonitoringResources object

NetworkConfig

The networking configuration for the data quality monitoring job.

Type: MonitoringNetworkConfig object

RoleArn

The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker AI can assume to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

StoppingCondition

A time limit for how long the monitoring job is allowed to run before stopping.

Type: MonitoringStoppingCondition object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeDevice

Service: Amazon SageMaker Service

Describes the device.

Request Syntax

```
{
   "DeviceFleetName": "string",
   "DeviceName": "string",
   "NextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeviceFleetName

The name of the fleet the devices belong to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

DeviceName

The unique ID of the device.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

NextToken

Next token of device description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "AgentVersion": "string",
   "Description": "string",
   "DeviceArn": "string",
   "DeviceFleetName": "string",
   "DeviceName": "string",
   "IotThingName": "string",
   "LatestHeartbeat": number,
   "MaxModels": number,
   "Models": [
      {
         "LatestInference": number,
         "LatestSampleTime": number,
         "ModelName": "string",
         "ModelVersion": "string"
      }
   ],
   "NextToken": "string",
   "RegistrationTime": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AgentVersion

Edge Manager agent version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 30.

Pattern: `[a-zA-Z0-9\ \_\.]+`

Description

A description of the device.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 40.

Pattern: `[-a-zA-Z0-9_.,;:! ]*`

DeviceArn

The Amazon Resource Name (ARN) of the device.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:[a-z\-]*:[a-z\-]*:\d{12}:[a-z\-]*/?[a-zA-Z_0-9+=,.@\-_/]+`

DeviceFleetName

The name of the fleet the device belongs to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

DeviceName

The unique identifier of the device.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

IotThingName

The AWS Internet of Things (IoT) object thing name associated with the device.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `[a-zA-Z0-9:_-]+`

LatestHeartbeat

The last heartbeat received from the device.

Type: Timestamp

MaxModels

The maximum number of models.

Type: Integer

Models

Models on the device.

Type: Array of EdgeModel objects

NextToken

The response from the last list when returning a list large enough to need tokening.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

RegistrationTime

The timestamp of the last registration or de-reregistration.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeDeviceFleet

Service: Amazon SageMaker Service

A description of the fleet the device belongs to.

Request Syntax

```
{
   "DeviceFleetName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeviceFleetName

The name of the fleet.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "Description": "string",
   "DeviceFleetArn": "string",
   "DeviceFleetName": "string",
   "IotRoleAlias": "string",
   "LastModifiedTime": number,
   "OutputConfig": {
      "KmsKeyId": "string",
      "PresetDeploymentConfig": "string",
```

```
      "PresetDeploymentType": "string",
      "S3OutputLocation": "string"
   },
   "RoleArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

Timestamp of when the device fleet was created.

Type: Timestamp

Description

A description of the fleet.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 800.

Pattern: `[-a-zA-Z0-9_.,;:! ]*`

DeviceFleetArn

The The Amazon Resource Name (ARN) of the fleet.

Type: String

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:device-fleet/?[a-zA-Z_0-9+=,.@\-_/]+`

DeviceFleetName

The name of the fleet.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

IotRoleAlias

The Amazon Resource Name (ARN) alias created in AWS Internet of Things (IoT).

Type: String

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:rolealias/?[a-zA-Z_0-9+=,.@\-_/]+`

LastModifiedTime

Timestamp of when the device fleet was last updated.

Type: Timestamp

OutputConfig

The output configuration for storing sampled data.

Type: EdgeOutputConfig object

RoleArn

The Amazon Resource Name (ARN) that has access to AWS Internet of Things (IoT).

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeDomain

Service: Amazon SageMaker Service

The description of the domain.

Request Syntax

```
{
   "DomainId": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

Response Syntax

```
{
   "AppNetworkAccessType": "string",
   "AppSecurityGroupManagement": "string",
   "AuthMode": "string",
   "CreationTime": number,
   "DefaultSpaceSettings": {
      "CustomFileSystemConfigs": [
         { ... }
      ],
      "CustomPosixUserConfig": {
         "Gid": number,
         "Uid": number
```

```
      },
      "ExecutionRole": "string",
      "JupyterLabAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number,
               "LifecycleManagement": "string",
               "MaxIdleTimeoutInMinutes": number,
               "MinIdleTimeoutInMinutes": number
            }
         },
         "BuiltInLifecycleConfigArn": "string",
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "EmrSettings": {
            "AssumableRoleArns": [ "string" ],
            "ExecutionRoleArns": [ "string" ]
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "JupyterServerAppSettings": {
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
```

```
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "KernelGatewayAppSettings": {
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "SecurityGroups": [ "string" ],
      "SpaceStorageSettings": {
         "DefaultEbsStorageSettings": {
            "DefaultEbsVolumeSizeInGb": number,
            "MaximumEbsVolumeSizeInGb": number
         }
      }
   },
   "DefaultUserSettings": {
      "AutoMountHomeEFS": "string",
      "CanvasAppSettings": {
         "DirectDeploySettings": {
            "Status": "string"
         },
         "EmrServerlessSettings": {
            "ExecutionRoleArn": "string",
```

```
            "Status": "string"
         },
         "GenerativeAiSettings": {
            "AmazonBedrockRoleArn": "string"
         },
         "IdentityProviderOAuthSettings": [
            {
               "DataSourceName": "string",
               "SecretArn": "string",
               "Status": "string"
            }
         ],
         "KendraSettings": {
            "Status": "string"
         },
         "ModelRegisterSettings": {
            "CrossAccountModelRegisterRoleArn": "string",
            "Status": "string"
         },
         "TimeSeriesForecastingSettings": {
            "AmazonForecastRoleArn": "string",
            "Status": "string"
         },
         "WorkspaceSettings": {
            "S3ArtifactPath": "string",
            "S3KmsKeyId": "string"
         }
      },
      "CodeEditorAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number,
               "LifecycleManagement": "string",
               "MaxIdleTimeoutInMinutes": number,
               "MinIdleTimeoutInMinutes": number
            }
         },
         "BuiltInLifecycleConfigArn": "string",
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
```

```
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "CustomFileSystemConfigs": [
         { ... }
      ],
      "CustomPosixUserConfig": {
         "Gid": number,
         "Uid": number
      },
      "DefaultLandingUri": "string",
      "ExecutionRole": "string",
      "JupyterLabAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number,
               "LifecycleManagement": "string",
               "MaxIdleTimeoutInMinutes": number,
               "MinIdleTimeoutInMinutes": number
            }
         },
         "BuiltInLifecycleConfigArn": "string",
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
```

```
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "EmrSettings": {
            "AssumableRoleArns": [ "string" ],
            "ExecutionRoleArns": [ "string" ]
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "JupyterServerAppSettings": {
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "KernelGatewayAppSettings": {
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
```

```
      "LifecycleConfigArns": [ "string" ]
   },
   "RSessionAppSettings": {
      "CustomImages": [
         {
            "AppImageConfigName": "string",
            "ImageName": "string",
            "ImageVersionNumber": number
         }
      ],
      "DefaultResourceSpec": {
         "InstanceType": "string",
         "LifecycleConfigArn": "string",
         "SageMakerImageArn": "string",
         "SageMakerImageVersionAlias": "string",
         "SageMakerImageVersionArn": "string",
         "TrainingPlanArn": "string"
      }
   },
   "RStudioServerProAppSettings": {
      "AccessStatus": "string",
      "UserGroup": "string"
   },
   "SecurityGroups": [ "string" ],
   "SharingSettings": {
      "NotebookOutputOption": "string",
      "S3KmsKeyId": "string",
      "S3OutputPath": "string"
   },
   "SpaceStorageSettings": {
      "DefaultEbsStorageSettings": {
         "DefaultEbsVolumeSizeInGb": number,
         "MaximumEbsVolumeSizeInGb": number
      }
   },
   "StudioWebPortal": "string",
   "StudioWebPortalSettings": {
      "ExecutionRoleSessionNameMode": "string",
      "HiddenAppTypes": [ "string" ],
      "HiddenInstanceTypes": [ "string" ],
      "HiddenMlTools": [ "string" ],
      "HiddenSageMakerImageVersionAliases": [
         {
            "SageMakerImageName": "string",
```

```
                "VersionAliases": [ "string" ]
            }
         ]
      },
      "TensorBoardAppSettings": {
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         }
      }
   },
   "DomainArn": "string",
   "DomainId": "string",
   "DomainName": "string",
   "DomainSettings": {
      "AmazonQSettings": {
         "QProfileArn": "string",
         "Status": "string"
      },
      "DockerSettings": {
         "EnableDockerAccess": "string",
         "RootlessDocker": "string",
         "VpcOnlyTrustedAccounts": [ "string" ]
      },
      "ExecutionRoleIdentityConfig": "string",
      "IpAddressType": "string",
      "RStudioServerProDomainSettings": {
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "DomainExecutionRoleArn": "string",
         "RStudioConnectUrl": "string",
         "RStudioPackageManagerUrl": "string"
      },
      "SecurityGroupIds": [ "string" ],
```

```
      "TrustedIdentityPropagationSettings": {
         "Status": "string"
      },
      "UnifiedStudioSettings": {
         "DomainAccountId": "string",
         "DomainId": "string",
         "DomainRegion": "string",
         "EnvironmentId": "string",
         "ProjectId": "string",
         "ProjectS3Path": "string",
         "SingleSignOnApplicationArn": "string",
         "StudioWebPortalAccess": "string"
      }
   },
   "FailureReason": "string",
   "HomeEfsFileSystemCreation": "string",
   "HomeEfsFileSystemId": "string",
   "HomeEfsFileSystemKmsKeyId": "string",
   "KmsKeyId": "string",
   "LastModifiedTime": number,
   "SecurityGroupIdForDomainBoundary": "string",
   "SingleSignOnApplicationArn": "string",
   "SingleSignOnManagedApplicationInstanceId": "string",
   "Status": "string",
   "SubnetIds": [ "string" ],
   "TagPropagation": "string",
   "Url": "string",
   "VpcId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AppNetworkAccessType

Specifies the VPC used for non-EFS traffic. The default value is `PublicInternetOnly`.

- `PublicInternetOnly` - Non-EFS traffic is through a VPC managed by Amazon SageMaker AI, which allows direct internet access
- `VpcOnly` - All traffic is through the specified VPC and subnets

Type: String

Valid Values: `PublicInternetOnly | VpcOnly`

AppSecurityGroupManagement

The entity that creates and manages the required security groups for inter-app communication in `VPCOnly` mode. Required when `CreateDomain.AppNetworkAccessType` is `VPCOnly` and `DomainSettings.RStudioServerProDomainSettings.DomainExecutionRoleArn` is provided.

Type: String

Valid Values: `Service | Customer`

AuthMode

The domain's authentication mode.

Type: String

Valid Values: `SSO | IAM`

CreationTime

The creation time.

Type: Timestamp

DefaultSpaceSettings

The default settings for shared spaces that users create in the domain.

Type: DefaultSpaceSettings object

DefaultUserSettings

Settings which are applied to UserProfiles in this domain if settings are not explicitly specified in a given UserProfile.

Type: UserSettings object

DomainArn

The domain's Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:domain/.*`

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

DomainName

The domain name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

DomainSettings

A collection of `Domain` settings.

Type: DomainSettings object

FailureReason

The failure reason.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

HomeEfsFileSystemCreation

Indicates whether a home EFS file system is created for the domain.

Type: String

Valid Values: `Enabled | Disabled`

HomeEfsFileSystemId

The ID of the Amazon Elastic File System managed by this Domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 32.

HomeEfsFileSystemKmsKeyId

This parameter has been deprecated.

Use `KmsKeyId`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

KmsKeyId

The AWS KMS customer managed key used to encrypt the EFS volume attached to the domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

LastModifiedTime

The last modified time.

Type: Timestamp

SecurityGroupIdForDomainBoundary

The ID of the security group that authorizes traffic between the `RSessionGateway` apps and the `RStudioServerPro` app.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

SingleSignOnApplicationArn

The ARN of the application managed by SageMaker AI in IAM Identity Center. This value is only returned for domains created after October 1, 2023.

Type: String

Pattern: `arn:(aws|aws-us-gov|aws-cn|aws-iso|aws-iso-b):sso::[0-9]+:application\/[a-zA-Z0-9-_.]+\/apl-[a-zA-Z0-9]+`

SingleSignOnManagedApplicationInstanceId

The IAM Identity Center managed application instance ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Status

The status.

Type: String

Valid Values: `Deleting | Failed | InService | Pending | Updating | Update_Failed | Delete_Failed`

SubnetIds

The VPC subnets that the domain uses for communication.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

TagPropagation

Indicates whether custom tag propagation is supported for the domain.

Type: String

Valid Values: `ENABLED | DISABLED`

Url

The domain's URL.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

VpcId

The ID of the Amazon Virtual Private Cloud (VPC) that the domain uses for communication.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeEdgeDeploymentPlan

Service: Amazon SageMaker Service

Describes an edge deployment plan with deployment status per stage.

Request Syntax

```
{
   "EdgeDeploymentPlanName": "string",
   "MaxResults": number,
   "NextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EdgeDeploymentPlanName

The name of the deployment plan to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

MaxResults

The maximum number of results to select (50 by default).

Type: Integer

Valid Range: Maximum value of 10.

Required: No

NextToken

If the edge deployment plan has enough stages to require tokening, then this is the response from the last list of stages returned.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "CreationTime": number,
   "DeviceFleetName": "string",
   "EdgeDeploymentFailed": number,
   "EdgeDeploymentPending": number,
   "EdgeDeploymentPlanArn": "string",
   "EdgeDeploymentPlanName": "string",
   "EdgeDeploymentSuccess": number,
   "LastModifiedTime": number,
   "ModelConfigs": [
      {
         "EdgePackagingJobName": "string",
         "ModelHandle": "string"
      }
   ],
   "NextToken": "string",
   "Stages": [
      {
         "DeploymentConfig": {
            "FailureHandlingPolicy": "string"
         },
         "DeploymentStatus": {
            "EdgeDeploymentFailedInStage": number,
            "EdgeDeploymentPendingInStage": number,
            "EdgeDeploymentStageStartTime": number,
            "EdgeDeploymentStatusMessage": "string",
            "EdgeDeploymentSuccessInStage": number,
            "StageStatus": "string"
         },
         "DeviceSelectionConfig": {
            "DeviceNameContains": "string",
            "DeviceNames": [ "string" ],
            "DeviceSubsetType": "string",
```

```
            "Percentage": number
         },
         "StageName": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The time when the edge deployment plan was created.

Type: Timestamp

DeviceFleetName

The device fleet used for this edge deployment plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

EdgeDeploymentFailed

The number of edge devices that failed the deployment.

Type: Integer

EdgeDeploymentPending

The number of edge devices yet to pick up deployment, or in progress.

Type: Integer

EdgeDeploymentPlanArn

The ARN of edge deployment plan.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z\-]*:\d{12}:edge-deployment/?[a-zA-Z_0-9+=,.@\-_/]+`

EdgeDeploymentPlanName

The name of the edge deployment plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

EdgeDeploymentSuccess

The number of edge devices with the successful deployment.

Type: Integer

LastModifiedTime

The time when the edge deployment plan was last updated.

Type: Timestamp

ModelConfigs

List of models associated with the edge deployment plan.

Type: Array of EdgeDeploymentModelConfig objects

NextToken

Token to use when calling the next set of stages in the edge deployment plan.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Stages

List of stages in the edge deployment plan.

Type: Array of DeploymentStageStatusSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeEdgePackagingJob

Service: Amazon SageMaker Service

A description of edge packaging jobs.

Request Syntax

```
{
   "EdgePackagingJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EdgePackagingJobName

The name of the edge packaging job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CompilationJobName": "string",
   "CreationTime": number,
   "EdgePackagingJobArn": "string",
   "EdgePackagingJobName": "string",
   "EdgePackagingJobStatus": "string",
   "EdgePackagingJobStatusMessage": "string",
   "LastModifiedTime": number,
   "ModelArtifact": "string",
   "ModelName": "string",
```

```
   "ModelSignature": "string",
   "ModelVersion": "string",
   "OutputConfig": {
      "KmsKeyId": "string",
      "PresetDeploymentConfig": "string",
      "PresetDeploymentType": "string",
      "S3OutputLocation": "string"
   },
   "PresetDeploymentOutput": {
      "Artifact": "string",
      "Status": "string",
      "StatusMessage": "string",
      "Type": "string"
   },
   "ResourceKey": "string",
   "RoleArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CompilationJobName

The name of the SageMaker Neo compilation job that is used to locate model artifacts that are being packaged.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

CreationTime

The timestamp of when the packaging job was created.

Type: Timestamp

EdgePackagingJobArn

The Amazon Resource Name (ARN) of the edge packaging job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z\-]*:\d{12}:edge-packaging-job/?[a-zA-Z_0-9+=,.@\-_/]+`

EdgePackagingJobName

The name of the edge packaging job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

EdgePackagingJobStatus

The current status of the packaging job.

Type: String

Valid Values: `STARTING | INPROGRESS | COMPLETED | FAILED | STOPPING | STOPPED`

EdgePackagingJobStatusMessage

Returns a message describing the job status and error messages.

Type: String

LastModifiedTime

The timestamp of when the job was last updated.

Type: Timestamp

ModelArtifact

The Amazon Simple Storage (S3) URI where model artifacts ares stored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

ModelName

The name of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ModelSignature

The signature document of files in the model artifact.

Type: String

ModelVersion

The version of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 30.

Pattern: `[a-zA-Z0-9\ \_\.]+`

OutputConfig

The output configuration for the edge packaging job.

Type: EdgeOutputConfig object

PresetDeploymentOutput

The output of a SageMaker Edge Manager deployable resource.

Type: EdgePresetDeploymentOutput object

ResourceKey

The AWS KMS key to use when encrypting the EBS volume the job run on.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

RoleArn

The Amazon Resource Name (ARN) of an IAM role that enables Amazon SageMaker to download and upload the model, and to contact Neo.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeEndpoint

Service: Amazon SageMaker Service

Returns the description of an endpoint.

Request Syntax

```
{
   "EndpointName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EndpointName

The name of the endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AsyncInferenceConfig": {
      "ClientConfig": {
         "MaxConcurrentInvocationsPerInstance": number
      },
      "OutputConfig": {
         "KmsKeyId": "string",
         "NotificationConfig": {
            "ErrorTopic": "string",
            "IncludeInferenceResponseIn": [ "string" ],
            "SuccessTopic": "string"
         },
```

```
            "S3FailurePath": "string",
            "S3OutputPath": "string"
         }
      },
      "CreationTime": number,
      "DataCaptureConfig": {
         "CaptureStatus": "string",
         "CurrentSamplingPercentage": number,
         "DestinationS3Uri": "string",
         "EnableCapture": boolean,
         "KmsKeyId": "string"
      },
      "EndpointArn": "string",
      "EndpointConfigName": "string",
      "EndpointName": "string",
      "EndpointStatus": "string",
      "ExplainerConfig": {
         "ClarifyExplainerConfig": {
            "EnableExplanations": "string",
            "InferenceConfig": {
               "ContentTemplate": "string",
               "FeatureHeaders": [ "string" ],
               "FeaturesAttribute": "string",
               "FeatureTypes": [ "string" ],
               "LabelAttribute": "string",
               "LabelHeaders": [ "string" ],
               "LabelIndex": number,
               "MaxPayloadInMB": number,
               "MaxRecordCount": number,
               "ProbabilityAttribute": "string",
               "ProbabilityIndex": number
            },
            "ShapConfig": {
               "NumberOfSamples": number,
               "Seed": number,
               "ShapBaselineConfig": {
                  "MimeType": "string",
                  "ShapBaseline": "string",
                  "ShapBaselineUri": "string"
               },
               "TextConfig": {
                  "Granularity": "string",
                  "Language": "string"
               },
```

```
            "UseLogit": boolean
         }
      }
   },
   "FailureReason": "string",
   "LastDeploymentConfig": {
      "AutoRollbackConfiguration": {
         "Alarms": [
            {
               "AlarmName": "string"
            }
         ]
      },
      "BlueGreenUpdatePolicy": {
         "MaximumExecutionTimeoutInSeconds": number,
         "TerminationWaitInSeconds": number,
         "TrafficRoutingConfiguration": {
            "CanarySize": {
               "Type": "string",
               "Value": number
            },
            "LinearStepSize": {
               "Type": "string",
               "Value": number
            },
            "Type": "string",
            "WaitIntervalInSeconds": number
         }
      },
      "RollingUpdatePolicy": {
         "MaximumBatchSize": {
            "Type": "string",
            "Value": number
         },
         "MaximumExecutionTimeoutInSeconds": number,
         "RollbackMaximumBatchSize": {
            "Type": "string",
            "Value": number
         },
         "WaitIntervalInSeconds": number
      }
   },
   "LastModifiedTime": number,
   "MetricsConfig": {
```

```
      "EnableDetailedObservability": boolean,
      "EnableEnhancedMetrics": boolean,
      "MetricPublishFrequencyInSeconds": number
   },
   "PendingDeploymentSummary": {
      "EndpointConfigName": "string",
      "ProductionVariants": [
         {
            "AcceleratorType": "string",
            "CurrentInstanceCount": number,
            "CurrentServerlessConfig": {
               "MaxConcurrency": number,
               "MemorySizeInMB": number,
               "ProvisionedConcurrency": number
            },
            "CurrentWeight": number,
            "DeployedImages": [
               {
                  "ResolutionTime": number,
                  "ResolvedImage": "string",
                  "SpecifiedImage": "string"
               }
            ],
            "DesiredInstanceCount": number,
            "DesiredServerlessConfig": {
               "MaxConcurrency": number,
               "MemorySizeInMB": number,
               "ProvisionedConcurrency": number
            },
            "DesiredWeight": number,
            "InstancePools": [
               {
                  "CurrentInstanceCount": number,
                  "InstanceType": "string"
               }
            ],
            "InstanceType": "string",
            "ManagedInstanceScaling": {
               "MaxInstanceCount": number,
               "MinInstanceCount": number,
               "ScaleInPolicy": {
                  "CooldownInMinutes": number,
                  "MaximumStepSize": number,
                  "Strategy": "string"
```

```
            },
            "Status": "string"
         },
         "RoutingConfig": {
            "RoutingStrategy": "string"
         },
         "VariantName": "string",
         "VariantStatus": [
            {
               "StartTime": number,
               "Status": "string",
               "StatusMessage": "string"
            }
         ]
      }
   ],
   "ShadowProductionVariants": [
      {
         "AcceleratorType": "string",
         "CurrentInstanceCount": number,
         "CurrentServerlessConfig": {
            "MaxConcurrency": number,
            "MemorySizeInMB": number,
            "ProvisionedConcurrency": number
         },
         "CurrentWeight": number,
         "DeployedImages": [
            {
               "ResolutionTime": number,
               "ResolvedImage": "string",
               "SpecifiedImage": "string"
            }
         ],
         "DesiredInstanceCount": number,
         "DesiredServerlessConfig": {
            "MaxConcurrency": number,
            "MemorySizeInMB": number,
            "ProvisionedConcurrency": number
         },
         "DesiredWeight": number,
         "InstancePools": [
            {
               "CurrentInstanceCount": number,
               "InstanceType": "string"
```

```
            }
         ],
         "InstanceType": "string",
         "ManagedInstanceScaling": {
            "MaxInstanceCount": number,
            "MinInstanceCount": number,
            "ScaleInPolicy": {
               "CooldownInMinutes": number,
               "MaximumStepSize": number,
               "Strategy": "string"
            },
            "Status": "string"
         },
         "RoutingConfig": {
            "RoutingStrategy": "string"
         },
         "VariantName": "string",
         "VariantStatus": [
            {
               "StartTime": number,
               "Status": "string",
               "StatusMessage": "string"
            }
         ]
      }
   ],
   "StartTime": number
},
"ProductionVariants": [
   {
      "CapacityReservationConfig": {
         "AvailableInstanceCount": number,
         "CapacityReservationPreference": "string",
         "Ec2CapacityReservations": [
            {
               "AvailableInstanceCount": number,
               "Ec2CapacityReservationId": "string",
               "TotalInstanceCount": number,
               "UsedByCurrentEndpoint": number
            }
         ],
         "MlReservationArn": "string",
         "TotalInstanceCount": number,
         "UsedByCurrentEndpoint": number
```

```
        },
        "CurrentInstanceCount": number,
        "CurrentServerlessConfig": {
           "MaxConcurrency": number,
           "MemorySizeInMB": number,
           "ProvisionedConcurrency": number
        },
        "CurrentWeight": number,
        "DeployedImages": [
           {
              "ResolutionTime": number,
              "ResolvedImage": "string",
              "SpecifiedImage": "string"
           }
        ],
        "DesiredInstanceCount": number,
        "DesiredServerlessConfig": {
           "MaxConcurrency": number,
           "MemorySizeInMB": number,
           "ProvisionedConcurrency": number
        },
        "DesiredWeight": number,
        "InstancePools": [
           {
              "CurrentInstanceCount": number,
              "InstanceType": "string"
           }
        ],
        "ManagedInstanceScaling": {
           "MaxInstanceCount": number,
           "MinInstanceCount": number,
           "ScaleInPolicy": {
              "CooldownInMinutes": number,
              "MaximumStepSize": number,
              "Strategy": "string"
           },
           "Status": "string"
        },
        "RoutingConfig": {
           "RoutingStrategy": "string"
        },
        "VariantName": "string",
        "VariantStatus": [
           {
```

```
            "StartTime": number,
            "Status": "string",
            "StatusMessage": "string"
         }
      ]
   }
],
"ShadowProductionVariants": [
   {
      "CapacityReservationConfig": {
         "AvailableInstanceCount": number,
         "CapacityReservationPreference": "string",
         "Ec2CapacityReservations": [
            {
               "AvailableInstanceCount": number,
               "Ec2CapacityReservationId": "string",
               "TotalInstanceCount": number,
               "UsedByCurrentEndpoint": number
            }
         ],
         "MlReservationArn": "string",
         "TotalInstanceCount": number,
         "UsedByCurrentEndpoint": number
      },
      "CurrentInstanceCount": number,
      "CurrentServerlessConfig": {
         "MaxConcurrency": number,
         "MemorySizeInMB": number,
         "ProvisionedConcurrency": number
      },
      "CurrentWeight": number,
      "DeployedImages": [
         {
            "ResolutionTime": number,
            "ResolvedImage": "string",
            "SpecifiedImage": "string"
         }
      ],
      "DesiredInstanceCount": number,
      "DesiredServerlessConfig": {
         "MaxConcurrency": number,
         "MemorySizeInMB": number,
         "ProvisionedConcurrency": number
      },
```

```
         "DesiredWeight": number,
         "InstancePools": [
            {
               "CurrentInstanceCount": number,
               "InstanceType": "string"
            }
         ],
         "ManagedInstanceScaling": {
            "MaxInstanceCount": number,
            "MinInstanceCount": number,
            "ScaleInPolicy": {
               "CooldownInMinutes": number,
               "MaximumStepSize": number,
               "Strategy": "string"
            },
            "Status": "string"
         },
         "RoutingConfig": {
            "RoutingStrategy": "string"
         },
         "VariantName": "string",
         "VariantStatus": [
            {
               "StartTime": number,
               "Status": "string",
               "StatusMessage": "string"
            }
         ]
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AsyncInferenceConfig

Returns the description of an endpoint configuration created using the `CreateEndpointConfig` API.

Type: AsyncInferenceConfig object

CreationTime

A timestamp that shows when the endpoint was created.

Type: Timestamp

DataCaptureConfig

The currently active data capture configuration used by your Endpoint.

Type: DataCaptureConfigSummary object

EndpointArn

The Amazon Resource Name (ARN) of the endpoint.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint/.*`

EndpointConfigName

The name of the endpoint configuration associated with this endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

EndpointName

Name of the endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

EndpointStatus

The status of the endpoint.

- `OutOfService`: Endpoint is not available to take incoming requests.

- `Creating`: CreateEndpoint is executing.
- `Updating`: UpdateEndpoint or UpdateEndpointWeightsAndCapacities is executing.
- `SystemUpdating`: Endpoint is undergoing maintenance and cannot be updated or deleted or re-scaled until it has completed. This maintenance operation does not change any customer-specified values such as VPC config, KMS encryption, model, instance type, or instance count.
- `RollingBack`: Endpoint fails to scale up or down or change its variant weight and is in the process of rolling back to its previous configuration. Once the rollback completes, endpoint returns to an `InService` status. This transitional status only applies to an endpoint that has autoscaling enabled and is undergoing variant weight or capacity changes as part of an UpdateEndpointWeightsAndCapacities call or when the UpdateEndpointWeightsAndCapacities operation is called explicitly.
- `InService`: Endpoint is available to process incoming requests.
- `Deleting`: DeleteEndpoint is executing.
- `Failed`: Endpoint could not be created, updated, or re-scaled. Use the `FailureReason` value returned by DescribeEndpoint for information about the failure. DeleteEndpoint is the only operation that can be performed on a failed endpoint.
- `UpdateRollbackFailed`: Both the rolling deployment and auto-rollback failed. Your endpoint is in service with a mix of the old and new endpoint configurations. For information about how to remedy this issue and restore the endpoint's status to `InService`, see Rolling Deployments.

Type: String

Valid Values: `OutOfService | Creating | Updating | SystemUpdating | RollingBack | InService | Deleting | Failed | UpdateRollbackFailed`

ExplainerConfig

The configuration parameters for an explainer.

Type: ExplainerConfig object

FailureReason

If the status of the endpoint is `Failed`, the reason why it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

LastDeploymentConfig

The most recent deployment configuration for the endpoint.

Type: DeploymentConfig object

LastModifiedTime

A timestamp that shows when the endpoint was last modified.

Type: Timestamp

MetricsConfig

The configuration parameters for utilization metrics.

Type: MetricsConfig object

PendingDeploymentSummary

Returns the summary of an in-progress deployment. This field is only returned when the endpoint is creating or updating with a new endpoint configuration.

Type: PendingDeploymentSummary object

ProductionVariants

An array of ProductionVariantSummary objects, one for each model hosted behind this endpoint.

Type: Array of ProductionVariantSummary objects

Array Members: Minimum number of 1 item.

ShadowProductionVariants

An array of ProductionVariantSummary objects, one for each model that you want to host at this endpoint in shadow mode with production traffic replicated from the model specified on `ProductionVariants`.

Type: Array of ProductionVariantSummary objects

Array Members: Minimum number of 1 item.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeEndpointConfig

Service: Amazon SageMaker Service

Returns the description of an endpoint configuration created using the `CreateEndpointConfig` API.

Request Syntax

```
{
   "EndpointConfigName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EndpointConfigName

The name of the endpoint configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AsyncInferenceConfig": {
      "ClientConfig": {
         "MaxConcurrentInvocationsPerInstance": number
      },
      "OutputConfig": {
         "KmsKeyId": "string",
         "NotificationConfig": {
            "ErrorTopic": "string",
            "IncludeInferenceResponseIn": [ "string" ],
```

```
            "SuccessTopic": "string"
         },
         "S3FailurePath": "string",
         "S3OutputPath": "string"
      }
   },
   "CreationTime": number,
   "DataCaptureConfig": {
      "CaptureContentTypeHeader": {
         "CsvContentTypes": [ "string" ],
         "JsonContentTypes": [ "string" ]
      },
      "CaptureOptions": [
         {
            "CaptureMode": "string"
         }
      ],
      "DestinationS3Uri": "string",
      "EnableCapture": boolean,
      "InitialSamplingPercentage": number,
      "KmsKeyId": "string"
   },
   "EnableNetworkIsolation": boolean,
   "EndpointConfigArn": "string",
   "EndpointConfigName": "string",
   "ExecutionRoleArn": "string",
   "ExplainerConfig": {
      "ClarifyExplainerConfig": {
         "EnableExplanations": "string",
         "InferenceConfig": {
            "ContentTemplate": "string",
            "FeatureHeaders": [ "string" ],
            "FeaturesAttribute": "string",
            "FeatureTypes": [ "string" ],
            "LabelAttribute": "string",
            "LabelHeaders": [ "string" ],
            "LabelIndex": number,
            "MaxPayloadInMB": number,
            "MaxRecordCount": number,
            "ProbabilityAttribute": "string",
            "ProbabilityIndex": number
         },
         "ShapConfig": {
            "NumberOfSamples": number,
```

```
            "Seed": number,
            "ShapBaselineConfig": {
               "MimeType": "string",
               "ShapBaseline": "string",
               "ShapBaselineUri": "string"
            },
            "TextConfig": {
               "Granularity": "string",
               "Language": "string"
            },
            "UseLogit": boolean
         }
      }
   },
   "KmsKeyId": "string",
   "MetricsConfig": {
      "EnableDetailedObservability": boolean,
      "EnableEnhancedMetrics": boolean,
      "MetricPublishFrequencyInSeconds": number
   },
   "ProductionVariants": [
      {
         "AcceleratorType": "string",
         "CapacityReservationConfig": {
            "CapacityReservationPreference": "string",
            "MlReservationArn": "string"
         },
         "ContainerStartupHealthCheckTimeoutInSeconds": number,
         "CoreDumpConfig": {
            "DestinationS3Uri": "string",
            "KmsKeyId": "string"
         },
         "EnableSSMAccess": boolean,
         "InferenceAmiVersion": "string",
         "InitialInstanceCount": number,
         "InitialVariantWeight": number,
         "InstancePools": [
            {
               "InstanceType": "string",
               "ModelNameOverride": "string",
               "Priority": number
            }
         ],
         "InstanceType": "string",
```

```
         "ManagedInstanceScaling": {
            "MaxInstanceCount": number,
            "MinInstanceCount": number,
            "ScaleInPolicy": {
               "CooldownInMinutes": number,
               "MaximumStepSize": number,
               "Strategy": "string"
            },
            "Status": "string"
         },
         "ModelDataDownloadTimeoutInSeconds": number,
         "ModelName": "string",
         "RoutingConfig": {
            "RoutingStrategy": "string"
         },
         "ServerlessConfig": {
            "MaxConcurrency": number,
            "MemorySizeInMB": number,
            "ProvisionedConcurrency": number
         },
         "VariantInstanceProvisionTimeoutInSeconds": number,
         "VariantName": "string",
         "VolumeSizeInGB": number
      }
   ],
   "ShadowProductionVariants": [
      {
         "AcceleratorType": "string",
         "CapacityReservationConfig": {
            "CapacityReservationPreference": "string",
            "MlReservationArn": "string"
         },
         "ContainerStartupHealthCheckTimeoutInSeconds": number,
         "CoreDumpConfig": {
            "DestinationS3Uri": "string",
            "KmsKeyId": "string"
         },
         "EnableSSMAccess": boolean,
         "InferenceAmiVersion": "string",
         "InitialInstanceCount": number,
         "InitialVariantWeight": number,
         "InstancePools": [
            {
               "InstanceType": "string",
```

```
            "ModelNameOverride": "string",
            "Priority": number
         }
      ],
      "InstanceType": "string",
      "ManagedInstanceScaling": {
         "MaxInstanceCount": number,
         "MinInstanceCount": number,
         "ScaleInPolicy": {
            "CooldownInMinutes": number,
            "MaximumStepSize": number,
            "Strategy": "string"
         },
         "Status": "string"
      },
      "ModelDataDownloadTimeoutInSeconds": number,
      "ModelName": "string",
      "RoutingConfig": {
         "RoutingStrategy": "string"
      },
      "ServerlessConfig": {
         "MaxConcurrency": number,
         "MemorySizeInMB": number,
         "ProvisionedConcurrency": number
      },
      "VariantInstanceProvisionTimeoutInSeconds": number,
      "VariantName": "string",
      "VolumeSizeInGB": number
   }
],
"VpcConfig": {
   "SecurityGroupIds": [ "string" ],
   "Subnets": [ "string" ]
}
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AsyncInferenceConfig

Returns the description of an endpoint configuration created using the `CreateEndpointConfig` API.

Type: AsyncInferenceConfig object

CreationTime

A timestamp that shows when the endpoint configuration was created.

Type: Timestamp

DataCaptureConfig

Configuration to control how SageMaker AI captures inference data.

Type: DataCaptureConfig object

EnableNetworkIsolation

Indicates whether all model containers deployed to the endpoint are isolated. If they are, no inbound or outbound network calls can be made to or from the model containers.

Type: Boolean

EndpointConfigArn

The Amazon Resource Name (ARN) of the endpoint configuration.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint-config/.*`

EndpointConfigName

Name of the SageMaker endpoint configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you assigned to the endpoint configuration.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

ExplainerConfig

The configuration parameters for an explainer.

Type: ExplainerConfig object

KmsKeyId

AWS KMS key ID Amazon SageMaker uses to encrypt data when storing it on the ML storage volume attached to the instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

MetricsConfig

The configuration parameters for utilization metrics.

Type: MetricsConfig object

ProductionVariants

An array of `ProductionVariant` objects, one for each model that you want to host at this endpoint.

Type: Array of ProductionVariant objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

ShadowProductionVariants

An array of `ProductionVariant` objects, one for each model that you want to host at this endpoint in shadow mode with production traffic replicated from the model specified on `ProductionVariants`.

Type: Array of ProductionVariant objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

VpcConfig

Specifies an Amazon Virtual Private Cloud (VPC) that your SageMaker jobs, hosted models, and compute resources have access to. You can control access to and from your resources by configuring a VPC. For more information, see Give SageMaker Access to Resources in your Amazon VPC.

Type: VpcConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeExperiment

Service: Amazon SageMaker Service

Provides a list of an experiment's properties.

Request Syntax

```
{
   "ExperimentName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ExperimentName

The name of the experiment to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
```

```
   "CreationTime": number,
   "Description": "string",
   "DisplayName": "string",
   "ExperimentArn": "string",
   "ExperimentName": "string",
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "Source": {
      "SourceArn": "string",
      "SourceType": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreatedBy

Who created the experiment.

Type: UserContext object

CreationTime

When the experiment was created.

Type: Timestamp

Description

The description of the experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

DisplayName

The name of the experiment as displayed. If `DisplayName` isn't specified, `ExperimentName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

ExperimentArn

The Amazon Resource Name (ARN) of the experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment/.*`

ExperimentName

The name of the experiment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

LastModifiedBy

Who last modified the experiment.

Type: UserContext object

LastModifiedTime

When the experiment was last modified.

Type: Timestamp

Source

The Amazon Resource Name (ARN) of the source and, optionally, the type.

Type: ExperimentSource object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeFeatureGroup

Service: Amazon SageMaker Service

Use this operation to describe a `FeatureGroup`. The response includes information on the creation time, `FeatureGroup` name, the unique identifier for each `FeatureGroup`, and more.

Request Syntax

```
{
   "FeatureGroupName": "string",
   "NextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

FeatureGroupName

The name or Amazon Resource Name (ARN) of the `FeatureGroup` you want described.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group\/)?([a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63})`

Required: Yes

NextToken

A token to resume pagination of the list of `Features` (`FeatureDefinitions`). 2,500 `Features` are returned by default.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "CreationTime": number,
   "Description": "string",
   "EventTimeFeatureName": "string",
   "FailureReason": "string",
   "FeatureDefinitions": [
      {
         "CollectionConfig": { ... },
         "CollectionType": "string",
         "FeatureName": "string",
         "FeatureType": "string"
      }
   ],
   "FeatureGroupArn": "string",
   "FeatureGroupName": "string",
   "FeatureGroupStatus": "string",
   "LastModifiedTime": number,
   "LastUpdateStatus": {
      "FailureReason": "string",
      "Status": "string"
   },
   "NextToken": "string",
   "OfflineStoreConfig": {
      "DataCatalogConfig": {
         "Catalog": "string",
         "Database": "string",
         "TableName": "string"
      },
      "DisableGlueTableCreation": boolean,
      "S3StorageConfig": {
         "KmsKeyId": "string",
         "ResolvedOutputS3Uri": "string",
         "S3Uri": "string"
      },
      "TableFormat": "string"
   },
   "OfflineStoreStatus": {
      "BlockedReason": "string",
      "Status": "string"
```

```
   },
   "OnlineStoreConfig": {
      "EnableOnlineStore": boolean,
      "SecurityConfig": {
         "KmsKeyId": "string"
      },
      "StorageType": "string",
      "TtlDuration": {
         "Unit": "string",
         "Value": number
      }
   },
   "OnlineStoreTotalSizeBytes": number,
   "RecordIdentifierFeatureName": "string",
   "RoleArn": "string",
   "ThroughputConfig": {
      "ProvisionedReadCapacityUnits": number,
      "ProvisionedWriteCapacityUnits": number,
      "ThroughputMode": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

A timestamp indicating when SageMaker created the `FeatureGroup`.

Type: Timestamp

Description

A free form description of the feature group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

EventTimeFeatureName

The name of the feature that stores the `EventTime` of a Record in a `FeatureGroup`.

An `EventTime` is a point in time when a new event occurs that corresponds to the creation or update of a `Record` in a `FeatureGroup`. All `Records` in the `FeatureGroup` have a corresponding `EventTime`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

FailureReason

The reason that the `FeatureGroup` failed to be replicated in the `OfflineStore`. This is failure can occur because:

- The `FeatureGroup` could not be created in the `OfflineStore`.
- The `FeatureGroup` could not be deleted from the `OfflineStore`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

FeatureDefinitions

A list of the `Features` in the `FeatureGroup`. Each feature is defined by a `FeatureName` and `FeatureType`.

Type: Array of FeatureDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 2500 items.

FeatureGroupArn

The Amazon Resource Name (ARN) of the `FeatureGroup`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group/.*`

FeatureGroupName

he name of the `FeatureGroup`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63}`

FeatureGroupStatus

The status of the feature group.

Type: String

Valid Values: `Creating | Created | CreateFailed | Deleting | DeleteFailed`

LastModifiedTime

A timestamp indicating when the feature group was last updated.

Type: Timestamp

LastUpdateStatus

A value indicating whether the update made to the feature group was successful.

Type: LastUpdateStatus object

NextToken

A token to resume pagination of the list of `Features` (`FeatureDefinitions`).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

OfflineStoreConfig

The configuration of the offline store. It includes the following configurations:

- Amazon S3 location of the offline store.
- Configuration of the Glue data catalog.
- Table format of the offline store.
- Option to disable the automatic creation of a Glue table for the offline store.
- Encryption configuration.

Type: OfflineStoreConfig object

OfflineStoreStatus

The status of the `OfflineStore`. Notifies you if replicating data into the `OfflineStore` has failed. Returns either: `Active` or `Blocked`

Type: OfflineStoreStatus object

OnlineStoreConfig

The configuration for the `OnlineStore`.

Type: OnlineStoreConfig object

OnlineStoreTotalSizeBytes

The size of the `OnlineStore` in bytes.

Type: Long

RecordIdentifierFeatureName

The name of the `Feature` used for `RecordIdentifier`, whose value uniquely identifies a record stored in the feature store.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

RoleArn

The Amazon Resource Name (ARN) of the IAM execution role used to persist data into the OfflineStore if an OfflineStoreConfig is provided.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

ThroughputConfig

Active throughput configuration of the feature group. There are two modes: `ON_DEMAND` and `PROVISIONED`. With on-demand mode, you are charged for data reads and writes that

your application performs on your feature group. You do not need to specify read and write throughput because Feature Store accommodates your workloads as they ramp up and down. You can switch a feature group to on-demand only once in a 24 hour period. With provisioned throughput mode, you specify the read and write capacity per second that you expect your application to require, and you are billed based on those limits. Exceeding provisioned throughput will result in your requests being throttled.

Note: `PROVISIONED` throughput mode is supported only for feature groups that are offline-only, or use the `Standard` tier online store.

Type: ThroughputConfigDescription object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeFeatureMetadata

Service: Amazon SageMaker Service

Shows the metadata for a feature within a feature group.

Request Syntax

```
{
   "FeatureGroupName": "string",
   "FeatureName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

FeatureGroupName

The name or Amazon Resource Name (ARN) of the feature group containing the feature.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group\/)?([a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63})`

Required: Yes

FeatureName

The name of the feature.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "Description": "string",
   "FeatureGroupArn": "string",
   "FeatureGroupName": "string",
   "FeatureName": "string",
   "FeatureType": "string",
   "LastModifiedTime": number,
   "Parameters": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

A timestamp indicating when the feature was created.

Type: Timestamp

Description

The description you added to describe the feature.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

FeatureGroupArn

The Amazon Resource Number (ARN) of the feature group that contains the feature.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group/.*`

FeatureGroupName

The name of the feature group that you've specified.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63}`

FeatureName

The name of the feature that you've specified.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

FeatureType

The data type of the feature.

Type: String

Valid Values: `Integral | Fractional | String`

LastModifiedTime

A timestamp indicating when the metadata for the feature group was modified. For example, if you add a parameter describing the feature, the timestamp changes to reflect the last time you

Type: Timestamp

Parameters

The key-value pairs that you added to describe the feature.

Type: Array of FeatureParameter objects

Array Members: Minimum number of 0 items. Maximum number of 25 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeFlowDefinition

Service: Amazon SageMaker Service

Returns information about the specified flow definition.

Request Syntax

```
{
   "FlowDefinitionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

FlowDefinitionName

The name of the flow definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "FailureReason": "string",
   "FlowDefinitionArn": "string",
   "FlowDefinitionName": "string",
   "FlowDefinitionStatus": "string",
   "HumanLoopActivationConfig": {
      "HumanLoopActivationConditionsConfig": {
         "HumanLoopActivationConditions": "string"
      }
   },
```

```
   "HumanLoopConfig": {
      "HumanTaskUiArn": "string",
      "PublicWorkforceTaskPrice": {
         "AmountInUsd": {
            "Cents": number,
            "Dollars": number,
            "TenthFractionsOfACent": number
         }
      },
      "TaskAvailabilityLifetimeInSeconds": number,
      "TaskCount": number,
      "TaskDescription": "string",
      "TaskKeywords": [ "string" ],
      "TaskTimeLimitInSeconds": number,
      "TaskTitle": "string",
      "WorkteamArn": "string"
   },
   "HumanLoopRequestSource": {
      "AwsManagedHumanLoopRequestSource": "string"
   },
   "OutputConfig": {
      "KmsKeyId": "string",
      "S3OutputPath": "string"
   },
   "RoleArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The timestamp when the flow definition was created.

Type: Timestamp

FailureReason

The reason your flow definition failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

FlowDefinitionArn

The Amazon Resource Name (ARN) of the flow defintion.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]+:[0-9]{12}:flow-definition/.*`

FlowDefinitionName

The Amazon Resource Name (ARN) of the flow definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9]){0,62}`

FlowDefinitionStatus

The status of the flow definition. Valid values are listed below.

Type: String

Valid Values: `Initializing | Active | Failed | Deleting`

HumanLoopActivationConfig

An object containing information about what triggers a human review workflow.

Type: HumanLoopActivationConfig object

HumanLoopConfig

An object containing information about who works on the task, the workforce task price, and other task details.

Type: HumanLoopConfig object

HumanLoopRequestSource

Container for configuring the source of human task requests. Used to specify if Amazon Rekognition or Amazon Textract is used as an integration source.

Type: HumanLoopRequestSource object

OutputConfig

An object containing information about the output file.

Type: FlowDefinitionOutputConfig object

RoleArn

The Amazon Resource Name (ARN) of the AWS Identity and Access Management (IAM) execution role for the flow definition.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeHub

Service: Amazon SageMaker Service

Describes a hub.

Request Syntax

```
{
   "HubName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HubName

The name of the hub to describe.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "FailureReason": "string",
   "HubArn": "string",
   "HubDescription": "string",
   "HubDisplayName": "string",
   "HubName": "string",
   "HubSearchKeywords": [ "string" ],
   "HubStatus": "string",
   "LastModifiedTime": number,
   "S3StorageConfig": {
```

```
        "S3OutputPath": "string"
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The date and time that the hub was created.

Type: Timestamp

FailureReason

The failure reason if importing hub content failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

HubArn

The Amazon Resource Name (ARN) of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

HubDescription

A description of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Pattern: `.*`

HubDisplayName

The display name of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

HubName

The name of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

HubSearchKeywords

The searchable keywords for the hub.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[^A-Z]*`

HubStatus

The status of the hub.

Type: String

Valid Values: `InService | Creating | Updating | Deleting | CreateFailed | UpdateFailed | DeleteFailed`

LastModifiedTime

The date and time that the hub was last modified.

Type: Timestamp

S3StorageConfig

The Amazon S3 storage configuration for the hub.

Type: HubS3StorageConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeHubContent

Service: Amazon SageMaker Service

Describe the content of a hub.

Request Syntax

```
{
   "HubContentName": "string",
   "HubContentType": "string",
   "HubContentVersion": "string",
   "HubName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HubContentName

The name of the content to describe.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubContentType

The type of content in the hub.

Type: String

Valid Values: `Model | Notebook | ModelReference | DataSet | JsonDoc`

Required: Yes

HubContentVersion

The version of the content to describe.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: No

HubName

The name of the hub that contains the content to describe.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "DocumentSchemaVersion": "string",
   "FailureReason": "string",
   "HubArn": "string",
   "HubContentArn": "string",
   "HubContentDependencies": [
      {
         "DependencyCopyPath": "string",
         "DependencyOriginPath": "string"
      }
   ],
   "HubContentDescription": "string",
   "HubContentDisplayName": "string",
   "HubContentDocument": "string",
   "HubContentMarkdown": "string",
   "HubContentName": "string",
   "HubContentSearchKeywords": [ "string" ],
   "HubContentStatus": "string",
   "HubContentType": "string",
   "HubContentVersion": "string",
   "HubName": "string",
```

```
   "LastModifiedTime": number,
   "ReferenceMinVersion": "string",
   "SageMakerPublicHubContentArn": "string",
   "SupportStatus": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The date and time that hub content was created.

Type: Timestamp

DocumentSchemaVersion

The document schema version for the hub content.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

FailureReason

The failure reason if importing hub content failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

HubArn

The Amazon Resource Name (ARN) of the hub that contains the content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

HubContentArn

The Amazon Resource Name (ARN) of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

HubContentDependencies

The location of any dependencies that the hub content has, such as scripts, model artifacts, datasets, or notebooks.

Type: Array of HubContentDependency objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

HubContentDescription

A description of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Pattern: `.*`

HubContentDisplayName

The display name of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

HubContentDocument

The hub content document that describes information about the hub content such as type, associated containers, scripts, and more.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 327680.

Pattern: `.*`

HubContentMarkdown

A string that provides a description of the hub content. This string can include links, tables, and standard markdown formating.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 170391.

HubContentName

The name of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

HubContentSearchKeywords

The searchable keywords for the hub content.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

HubContentStatus

The status of the hub content.

Type: String

Valid Values: `Available | Importing | Deleting | ImportFailed | DeleteFailed | PendingImport | PendingDelete`

HubContentType

The type of hub content.

Type: String

Valid Values: `Model | Notebook | ModelReference | DataSet | JsonDoc`

HubContentVersion

The version of the hub content.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

HubName

The name of the hub that contains the content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

LastModifiedTime

The last modified time of the hub content.

Type: Timestamp

ReferenceMinVersion

The minimum version of the hub content.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

SageMakerPublicHubContentArn

The ARN of the public hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:aws:hub-content\/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}\/Model\/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

SupportStatus

The support status of the hub content.

Type: String

Valid Values: `Supported | Deprecated | Restricted`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeHumanTaskUi

Service: Amazon SageMaker Service

Returns information about the requested human task user interface (worker task template).

Request Syntax

```
{
   "HumanTaskUiName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HumanTaskUiName

The name of the human task user interface (worker task template) you want information about.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9])*`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "HumanTaskUiArn": "string",
   "HumanTaskUiName": "string",
   "HumanTaskUiStatus": "string",
   "UiTemplate": {
      "ContentSha256": "string",
      "Url": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The timestamp when the human task user interface was created.

Type: Timestamp

HumanTaskUiArn

The Amazon Resource Name (ARN) of the human task user interface (worker task template).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]+:[0-9]{12}:human-task-ui/.*`

HumanTaskUiName

The name of the human task user interface (worker task template).

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9])*`

HumanTaskUiStatus

The status of the human task user interface (worker task template). Valid values are listed below.

Type: String

Valid Values: `Active | Deleting`

UiTemplate

Container for user interface template information.

Type: UiTemplateInfo object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeHyperParameterTuningJob

Service: Amazon SageMaker Service

Returns a description of a hyperparameter tuning job, depending on the fields selected. These fields can include the name, Amazon Resource Name (ARN), job status of your tuning job and more.

Request Syntax

```
{
   "HyperParameterTuningJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HyperParameterTuningJobName

The name of the tuning job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

Response Syntax

```
{
   "Autotune": {
      "Mode": "string"
   },
   "BestTrainingJob": {
      "CreationTime": number,
      "FailureReason": "string",
      "FinalHyperParameterTuningJobObjectiveMetric": {
         "MetricName": "string",
         "Type": "string",
```

```
            "Value": number
         },
         "ObjectiveStatus": "string",
         "TrainingEndTime": number,
         "TrainingJobArn": "string",
         "TrainingJobDefinitionName": "string",
         "TrainingJobName": "string",
         "TrainingJobStatus": "string",
         "TrainingStartTime": number,
         "TunedHyperParameters": {
            "string" : "string"
         },
         "TuningJobName": "string"
      },
      "ConsumedResources": {
         "RuntimeInSeconds": number
      },
      "CreationTime": number,
      "FailureReason": "string",
      "HyperParameterTuningEndTime": number,
      "HyperParameterTuningJobArn": "string",
      "HyperParameterTuningJobConfig": {
         "HyperParameterTuningJobObjective": {
            "MetricName": "string",
            "Type": "string"
         },
         "ParameterRanges": {
            "AutoParameters": [
               {
                  "Name": "string",
                  "ValueHint": "string"
               }
            ],
            "CategoricalParameterRanges": [
               {
                  "Name": "string",
                  "Values": [ "string" ]
               }
            ],
            "ContinuousParameterRanges": [
               {
                  "MaxValue": "string",
                  "MinValue": "string",
                  "Name": "string",
```

```
            "ScalingType": "string"
         }
      ],
      "IntegerParameterRanges": [
         {
            "MaxValue": "string",
            "MinValue": "string",
            "Name": "string",
            "ScalingType": "string"
         }
      ]
   },
   "RandomSeed": number,
   "ResourceLimits": {
      "MaxNumberOfTrainingJobs": number,
      "MaxParallelTrainingJobs": number,
      "MaxRuntimeInSeconds": number
   },
   "Strategy": "string",
   "StrategyConfig": {
      "HyperbandStrategyConfig": {
         "MaxResource": number,
         "MinResource": number
      }
   },
   "TrainingJobEarlyStoppingType": "string",
   "TuningJobCompletionCriteria": {
      "BestObjectiveNotImproving": {
         "MaxNumberOfTrainingJobsNotImproving": number
      },
      "ConvergenceDetected": {
         "CompleteOnConvergence": "string"
      },
      "TargetObjectiveMetricValue": number
   }
},
"HyperParameterTuningJobName": "string",
"HyperParameterTuningJobStatus": "string",
"LastModifiedTime": number,
"ObjectiveStatusCounters": {
   "Failed": number,
   "Pending": number,
   "Succeeded": number
},
```

```
   "OverallBestTrainingJob": {
      "CreationTime": number,
      "FailureReason": "string",
      "FinalHyperParameterTuningJobObjectiveMetric": {
         "MetricName": "string",
         "Type": "string",
         "Value": number
      },
      "ObjectiveStatus": "string",
      "TrainingEndTime": number,
      "TrainingJobArn": "string",
      "TrainingJobDefinitionName": "string",
      "TrainingJobName": "string",
      "TrainingJobStatus": "string",
      "TrainingStartTime": number,
      "TunedHyperParameters": {
         "string" : "string"
      },
      "TuningJobName": "string"
   },
   "TrainingJobDefinition": {
      "AlgorithmSpecification": {
         "AlgorithmName": "string",
         "MetricDefinitions": [
            {
               "Name": "string",
               "Regex": "string"
            }
         ],
         "TrainingImage": "string",
         "TrainingInputMode": "string"
      },
      "CheckpointConfig": {
         "LocalPath": "string",
         "S3Uri": "string"
      },
      "DefinitionName": "string",
      "EnableInterContainerTrafficEncryption": boolean,
      "EnableManagedSpotTraining": boolean,
      "EnableNetworkIsolation": boolean,
      "Environment": {
         "string" : "string"
      },
      "HyperParameterRanges": {
```

```
         "AutoParameters": [
            {
               "Name": "string",
               "ValueHint": "string"
            }
         ],
         "CategoricalParameterRanges": [
            {
               "Name": "string",
               "Values": [ "string" ]
            }
         ],
         "ContinuousParameterRanges": [
            {
               "MaxValue": "string",
               "MinValue": "string",
               "Name": "string",
               "ScalingType": "string"
            }
         ],
         "IntegerParameterRanges": [
            {
               "MaxValue": "string",
               "MinValue": "string",
               "Name": "string",
               "ScalingType": "string"
            }
         ]
      },
      "HyperParameterTuningResourceConfig": {
         "AllocationStrategy": "string",
         "InstanceConfigs": [
            {
               "InstanceCount": number,
               "InstanceType": "string",
               "VolumeSizeInGB": number
            }
         ],
         "InstanceCount": number,
         "InstanceType": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
      },
      "InputDataConfig": [
```

```
        {
           "ChannelName": "string",
           "CompressionType": "string",
           "ContentType": "string",
           "DataSource": {
              "DatasetSource": {
                 "DatasetArn": "string"
              },
              "FileSystemDataSource": {
                 "DirectoryPath": "string",
                 "FileSystemAccessMode": "string",
                 "FileSystemId": "string",
                 "FileSystemType": "string"
              },
              "S3DataSource": {
                 "AttributeNames": [ "string" ],
                 "HubAccessConfig": {
                    "HubContentArn": "string"
                 },
                 "InstanceGroupNames": [ "string" ],
                 "ModelAccessConfig": {
                    "AcceptEula": boolean
                 },
                 "S3DataDistributionType": "string",
                 "S3DataType": "string",
                 "S3Uri": "string"
              }
           },
           "InputMode": "string",
           "RecordWrapperType": "string",
           "ShuffleConfig": {
              "Seed": number
           }
        }
     ],
     "OutputDataConfig": {
        "CompressionType": "string",
        "KmsKeyId": "string",
        "S3OutputPath": "string"
     },
     "ResourceConfig": {
        "InstanceCount": number,
        "InstanceGroups": [
           {
```

```
            "InstanceCount": number,
            "InstanceGroupName": "string",
            "InstanceType": "string"
         }
      ],
      "InstancePlacementConfig": {
         "EnableMultipleJobs": boolean,
         "PlacementSpecifications": [
            {
               "InstanceCount": number,
               "UltraServerId": "string"
            }
         ]
      },
      "InstanceType": "string",
      "KeepAlivePeriodInSeconds": number,
      "TrainingPlanArn": "string",
      "VolumeKmsKeyId": "string",
      "VolumeSizeInGB": number
   },
   "RetryStrategy": {
      "MaximumRetryAttempts": number
   },
   "RoleArn": "string",
   "StaticHyperParameters": {
      "string" : "string"
   },
   "StoppingCondition": {
      "MaxPendingTimeInSeconds": number,
      "MaxRuntimeInSeconds": number,
      "MaxWaitTimeInSeconds": number
   },
   "TuningObjective": {
      "MetricName": "string",
      "Type": "string"
   },
   "VpcConfig": {
      "SecurityGroupIds": [ "string" ],
      "Subnets": [ "string" ]
   }
},
"TrainingJobDefinitions": [
   {
      "AlgorithmSpecification": {
```

```
            "AlgorithmName": "string",
            "MetricDefinitions": [
               {
                  "Name": "string",
                  "Regex": "string"
               }
            ],
            "TrainingImage": "string",
            "TrainingInputMode": "string"
         },
         "CheckpointConfig": {
            "LocalPath": "string",
            "S3Uri": "string"
         },
         "DefinitionName": "string",
         "EnableInterContainerTrafficEncryption": boolean,
         "EnableManagedSpotTraining": boolean,
         "EnableNetworkIsolation": boolean,
         "Environment": {
            "string" : "string"
         },
         "HyperParameterRanges": {
            "AutoParameters": [
               {
                  "Name": "string",
                  "ValueHint": "string"
               }
            ],
            "CategoricalParameterRanges": [
               {
                  "Name": "string",
                  "Values": [ "string" ]
               }
            ],
            "ContinuousParameterRanges": [
               {
                  "MaxValue": "string",
                  "MinValue": "string",
                  "Name": "string",
                  "ScalingType": "string"
               }
            ],
            "IntegerParameterRanges": [
               {
```

```
                "MaxValue": "string",
                "MinValue": "string",
                "Name": "string",
                "ScalingType": "string"
             }
          ]
       },
       "HyperParameterTuningResourceConfig": {
          "AllocationStrategy": "string",
          "InstanceConfigs": [
             {
                "InstanceCount": number,
                "InstanceType": "string",
                "VolumeSizeInGB": number
             }
          ],
          "InstanceCount": number,
          "InstanceType": "string",
          "VolumeKmsKeyId": "string",
          "VolumeSizeInGB": number
       },
       "InputDataConfig": [
          {
             "ChannelName": "string",
             "CompressionType": "string",
             "ContentType": "string",
             "DataSource": {
                "DatasetSource": {
                   "DatasetArn": "string"
                },
                "FileSystemDataSource": {
                   "DirectoryPath": "string",
                   "FileSystemAccessMode": "string",
                   "FileSystemId": "string",
                   "FileSystemType": "string"
                },
                "S3DataSource": {
                   "AttributeNames": [ "string" ],
                   "HubAccessConfig": {
                      "HubContentArn": "string"
                   },
                   "InstanceGroupNames": [ "string" ],
                   "ModelAccessConfig": {
                      "AcceptEula": boolean
```

```
                    },
                    "S3DataDistributionType": "string",
                    "S3DataType": "string",
                    "S3Uri": "string"
                 }
              },
              "InputMode": "string",
              "RecordWrapperType": "string",
              "ShuffleConfig": {
                 "Seed": number
              }
           }
        ],
        "OutputDataConfig": {
           "CompressionType": "string",
           "KmsKeyId": "string",
           "S3OutputPath": "string"
        },
        "ResourceConfig": {
           "InstanceCount": number,
           "InstanceGroups": [
              {
                 "InstanceCount": number,
                 "InstanceGroupName": "string",
                 "InstanceType": "string"
              }
           ],
           "InstancePlacementConfig": {
              "EnableMultipleJobs": boolean,
              "PlacementSpecifications": [
                 {
                    "InstanceCount": number,
                    "UltraServerId": "string"
                 }
              ]
           },
           "InstanceType": "string",
           "KeepAlivePeriodInSeconds": number,
           "TrainingPlanArn": "string",
           "VolumeKmsKeyId": "string",
           "VolumeSizeInGB": number
        },
        "RetryStrategy": {
           "MaximumRetryAttempts": number
```

```
         },
         "RoleArn": "string",
         "StaticHyperParameters": {
            "string" : "string"
         },
         "StoppingCondition": {
            "MaxPendingTimeInSeconds": number,
            "MaxRuntimeInSeconds": number,
            "MaxWaitTimeInSeconds": number
         },
         "TuningObjective": {
            "MetricName": "string",
            "Type": "string"
         },
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      }
   ],
   "TrainingJobStatusCounters": {
      "Completed": number,
      "InProgress": number,
      "NonRetryableError": number,
      "RetryableError": number,
      "Stopped": number
   },
   "TuningJobCompletionDetails": {
      "ConvergenceDetectedTime": number,
      "NumberOfTrainingJobsObjectiveNotImproving": number
   },
   "WarmStartConfig": {
      "ParentHyperParameterTuningJobs": [
         {
            "HyperParameterTuningJobName": "string"
         }
      ],
      "WarmStartType": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Autotune

A flag to indicate if autotune is enabled for the hyperparameter tuning job.

Type: Autotune object

BestTrainingJob

A TrainingJobSummary object that describes the training job that completed with the best current HyperParameterTuningJobObjective.

Type: HyperParameterTrainingJobSummary object

ConsumedResources

The total resources consumed by your hyperparameter tuning job.

Type: HyperParameterTuningJobConsumedResources object

CreationTime

The date and time that the tuning job started.

Type: Timestamp

FailureReason

If the tuning job failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

HyperParameterTuningEndTime

The date and time that the tuning job ended.

Type: Timestamp

HyperParameterTuningJobArn

The Amazon Resource Name (ARN) of the tuning job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:hyper-parameter-tuning-job/.*`

HyperParameterTuningJobConfig

The HyperParameterTuningJobConfig object that specifies the configuration of the tuning job.

Type: HyperParameterTuningJobConfig object

HyperParameterTuningJobName

The name of the hyperparameter tuning job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

HyperParameterTuningJobStatus

The status of the tuning job.

Type: String

Valid Values: `Completed | InProgress | Failed | Stopped | Stopping | Deleting | DeleteFailed`

LastModifiedTime

The date and time that the status of the tuning job was modified.

Type: Timestamp

ObjectiveStatusCounters

The ObjectiveStatusCounters object that specifies the number of training jobs, categorized by the status of their final objective metric, that this tuning job launched.

Type: ObjectiveStatusCounters object

OverallBestTrainingJob

If the hyperparameter tuning job is an warm start tuning job with a `WarmStartType` of `IDENTICAL_DATA_AND_ALGORITHM`, this is the TrainingJobSummary for the training job with the best objective metric value of all training jobs launched by this tuning job and all parent jobs specified for the warm start tuning job.

Type: HyperParameterTrainingJobSummary object

TrainingJobDefinition

The HyperParameterTrainingJobDefinition object that specifies the definition of the training jobs that this tuning job launches.

Type: HyperParameterTrainingJobDefinition object

TrainingJobDefinitions

A list of the HyperParameterTrainingJobDefinition objects launched for this tuning job.

Type: Array of HyperParameterTrainingJobDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

TrainingJobStatusCounters

The TrainingJobStatusCounters object that specifies the number of training jobs, categorized by status, that this tuning job launched.

Type: TrainingJobStatusCounters object

TuningJobCompletionDetails

Tuning job completion information returned as the response from a hyperparameter tuning job. This information tells if your tuning job has or has not converged. It also includes the number of training jobs that have not improved model performance as evaluated against the objective function.

Type: HyperParameterTuningJobCompletionDetails object

WarmStartConfig

The configuration for starting the hyperparameter parameter tuning job using one or more previous tuning jobs as a starting point. The results of previous tuning jobs are used to inform which combinations of hyperparameters to search over in the new tuning job.

Type: HyperParameterTuningJobWarmStartConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeImage

Service: Amazon SageMaker Service

Describes a SageMaker AI image.

Request Syntax

```
{
   "ImageName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ImageName

The name of the image to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "Description": "string",
   "DisplayName": "string",
   "FailureReason": "string",
   "ImageArn": "string",
   "ImageName": "string",
   "ImageStatus": "string",
   "LastModifiedTime": number,
   "RoleArn": "string"
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

When the image was created.

Type: Timestamp

Description

The description of the image.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 512.

Pattern: `.*`

DisplayName

The name of the image as displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `\S(.*\S)?`

FailureReason

When a create, update, or delete operation fails, the reason for the failure.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

ImageArn

The ARN of the image.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image/[a-zA-Z0-9]([-.]?[a-zA-Z0-9])*`

ImageName

The name of the image.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

ImageStatus

The status of the image.

Type: String

Valid Values: `CREATING | CREATED | CREATE_FAILED | UPDATING | UPDATE_FAILED | DELETING | DELETE_FAILED`

LastModifiedTime

When the image was last modified.

Type: Timestamp

RoleArn

The ARN of the IAM role that enables Amazon SageMaker AI to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeImageVersion

Service: Amazon SageMaker Service

Describes a version of a SageMaker AI image.

Request Syntax

```
{
   "Alias": "string",
   "ImageName": "string",
   "Version": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Alias

The alias of the image version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(?!^[.-])^([a-zA-Z0-9-_.]+)`

Required: No

ImageName

The name of the image.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

Version

The version of the image. If not specified, the latest version is described.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

Response Syntax

```
{
   "BaseImage": "string",
   "ContainerImage": "string",
   "CreationTime": number,
   "FailureReason": "string",
   "Horovod": boolean,
   "ImageArn": "string",
   "ImageVersionArn": "string",
   "ImageVersionStatus": "string",
   "JobType": "string",
   "LastModifiedTime": number,
   "MLFramework": "string",
   "Processor": "string",
   "ProgrammingLang": "string",
   "ReleaseNotes": "string",
   "VendorGuidance": "string",
   "Version": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

BaseImage

The registry path of the container image on which this image version is based.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.*`

ContainerImage

The registry path of the container image that contains this image version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

CreationTime

When the version was created.

Type: Timestamp

FailureReason

When a create or delete operation fails, the reason for the failure.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Horovod

Indicates Horovod compatibility.

Type: Boolean

ImageArn

The ARN of the image the version is based on.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image/[a-zA-Z0-9]([-.]?[a-zA-Z0-9])*`

ImageVersionArn

The ARN of the version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image-version/[a-z0-9]([-.]?[a-z0-9])*/[0-9]+|None)`

ImageVersionStatus

The status of the version.

Type: String

Valid Values: `CREATING | CREATED | CREATE_FAILED | DELETING | DELETE_FAILED`

JobType

Indicates SageMaker AI job type compatibility.

- `TRAINING`: The image version is compatible with SageMaker AI training jobs.
- `INFERENCE`: The image version is compatible with SageMaker AI inference jobs.
- `NOTEBOOK_KERNEL`: The image version is compatible with SageMaker AI notebook kernels.

Type: String

Valid Values: `TRAINING | INFERENCE | NOTEBOOK_KERNEL`

LastModifiedTime

When the version was last modified.

Type: Timestamp

MLFramework

The machine learning framework vended in the image version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z]+ ?\d+\.\d+(\.\d+)?`

Processor

Indicates CPU or GPU compatibility.

- `CPU`: The image version is compatible with CPU.

- GPU: The image version is compatible with GPU.

Type: String

Valid Values: CPU | GPU

ProgrammingLang

The supported programming language and its version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: [a-zA-Z]+ ?\d+\.\d+(\.\d+)?

ReleaseNotes

The maintainer description of the image version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: .*

VendorGuidance

The stability of the image version specified by the maintainer.

- NOT_PROVIDED: The maintainers did not provide a status for image version stability.
- STABLE: The image version is stable.
- TO_BE_ARCHIVED: The image version is set to be archived. Custom image versions that are set to be archived are automatically archived after three months.
- ARCHIVED: The image version is archived. Archived image versions are not searchable and are no longer actively supported.

Type: String

Valid Values: NOT_PROVIDED | STABLE | TO_BE_ARCHIVED | ARCHIVED

Version

The version number.

Type: Integer

Valid Range: Minimum value of 0.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeInferenceComponent

Service: Amazon SageMaker Service

Returns information about an inference component.

Request Syntax

```
{
   "InferenceComponentName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

InferenceComponentName

The name of the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "EndpointArn": "string",
   "EndpointName": "string",
   "FailureReason": "string",
   "InferenceComponentArn": "string",
   "InferenceComponentName": "string",
   "InferenceComponentStatus": "string",
   "LastDeploymentConfig": {
      "AutoRollbackConfiguration": {
         "Alarms": [
            {
```

```
            "AlarmName": "string"
         }
      ]
   },
   "RollingUpdatePolicy": {
      "MaximumBatchSize": {
         "Type": "string",
         "Value": number
      },
      "MaximumExecutionTimeoutInSeconds": number,
      "RollbackMaximumBatchSize": {
         "Type": "string",
         "Value": number
      },
      "WaitIntervalInSeconds": number
   }
},
"LastModifiedTime": number,
"RuntimeConfig": {
   "CurrentCopyCount": number,
   "DesiredCopyCount": number,
   "PlacementStatus": [
      {
         "CurrentCopyCount": number,
         "InstanceType": "string"
      }
   ]
},
"Specification": {
   "BaseInferenceComponentName": "string",
   "ComputeResourceRequirements": {
      "MaxMemoryRequiredInMb": number,
      "MinMemoryRequiredInMb": number,
      "NumberOfAcceleratorDevicesRequired": number,
      "NumberOfCpuCoresRequired": number
   },
   "Container": {
      "ArtifactUrl": "string",
      "ContainerMetricsConfig": {
         "MetricsEndpoints": [
            {
               "MetricPublishFrequencyInSeconds": number,
               "MetricsEndpointPath": "string"
            }
```

```
            ]
         },
         "DeployedImage": {
            "ResolutionTime": number,
            "ResolvedImage": "string",
            "SpecifiedImage": "string"
         },
         "Environment": {
            "string" : "string"
         }
      },
      "DataCacheConfig": {
         "EnableCaching": boolean
      },
      "InstanceType": "string",
      "ModelName": "string",
      "SchedulingConfig": {
         "AvailabilityZoneBalance": {
            "EnforcementMode": "string",
            "MaxImbalance": number
         },
         "PlacementStrategy": "string"
      },
      "StartupParameters": {
         "ContainerStartupHealthCheckTimeoutInSeconds": number,
         "ModelDataDownloadTimeoutInSeconds": number
      }
   },
   "Specifications": [
      {
         "BaseInferenceComponentName": "string",
         "ComputeResourceRequirements": {
            "MaxMemoryRequiredInMb": number,
            "MinMemoryRequiredInMb": number,
            "NumberOfAcceleratorDevicesRequired": number,
            "NumberOfCpuCoresRequired": number
         },
         "Container": {
            "ArtifactUrl": "string",
            "ContainerMetricsConfig": {
               "MetricsEndpoints": [
                  {
                     "MetricPublishFrequencyInSeconds": number,
                     "MetricsEndpointPath": "string"
```

```
                    }
                ]
            },
            "DeployedImage": {
                "ResolutionTime": number,
                "ResolvedImage": "string",
                "SpecifiedImage": "string"
            },
            "Environment": {
                "string" : "string"
            }
        },
        "DataCacheConfig": {
            "EnableCaching": boolean
        },
        "InstanceType": "string",
        "ModelName": "string",
        "SchedulingConfig": {
            "AvailabilityZoneBalance": {
                "EnforcementMode": "string",
                "MaxImbalance": number
            },
            "PlacementStrategy": "string"
        },
        "StartupParameters": {
            "ContainerStartupHealthCheckTimeoutInSeconds": number,
            "ModelDataDownloadTimeoutInSeconds": number
        }
    }
  ],
  "VariantName": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The time when the inference component was created.

Type: Timestamp

EndpointArn

The Amazon Resource Name (ARN) of the endpoint that hosts the inference component.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint/.*`

EndpointName

The name of the endpoint that hosts the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

FailureReason

If the inference component status is `Failed`, the reason for the failure.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

InferenceComponentArn

The Amazon Resource Name (ARN) of the inference component.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

InferenceComponentName

The name of the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

InferenceComponentStatus

The status of the inference component.

Type: String

Valid Values: `InService | Creating | Updating | Failed | Deleting`

LastDeploymentConfig

The deployment and rollback settings that you assigned to the inference component.

Type: InferenceComponentDeploymentConfig object

LastModifiedTime

The time when the inference component was last updated.

Type: Timestamp

RuntimeConfig

Details about the runtime settings for the model that is deployed with the inference component.

Type: InferenceComponentRuntimeConfigSummary object

Specification

Details about the resources that are deployed with this inference component.

Type: InferenceComponentSpecificationSummary object

Specifications

A list of specification summaries for the inference component, one per instance type. This parameter is populated when the inference component was created with multiple specifications. When this parameter is populated, the singular `Specification` parameter is not returned.

Type: Array of InferenceComponentSpecificationSummary objects

Array Members: Minimum number of 1 item.

VariantName

The name of the production variant that hosts the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeInferenceExperiment

Service: Amazon SageMaker Service

Returns details about an inference experiment.

Request Syntax

```
{
   "Name": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Name

The name of the inference experiment to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "Arn": "string",
   "CompletionTime": number,
   "CreationTime": number,
   "DataStorageConfig": {
      "ContentType": {
         "CsvContentTypes": [ "string" ],
         "JsonContentTypes": [ "string" ]
      },
      "Destination": "string",
      "KmsKey": "string"
   },
```

```
   "Description": "string",
   "EndpointMetadata": {
      "EndpointConfigName": "string",
      "EndpointName": "string",
      "EndpointStatus": "string",
      "FailureReason": "string"
   },
   "KmsKey": "string",
   "LastModifiedTime": number,
   "ModelVariants": [
      {
         "InfrastructureConfig": {
            "InfrastructureType": "string",
            "RealTimeInferenceConfig": {
               "InstanceCount": number,
               "InstanceType": "string"
            }
         },
         "ModelName": "string",
         "Status": "string",
         "VariantName": "string"
      }
   ],
   "Name": "string",
   "RoleArn": "string",
   "Schedule": {
      "EndTime": number,
      "StartTime": number
   },
   "ShadowModeConfig": {
      "ShadowModelVariants": [
         {
            "SamplingPercentage": number,
            "ShadowModelVariantName": "string"
         }
      ],
      "SourceModelVariantName": "string"
   },
   "Status": "string",
   "StatusReason": "string",
   "Type": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The ARN of the inference experiment being described.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:inference-experiment/.*`

CompletionTime

The timestamp at which the inference experiment was completed.

Type: Timestamp

CreationTime

The timestamp at which you created the inference experiment.

Type: Timestamp

DataStorageConfig

The Amazon S3 location and configuration for storing inference request and response data.

Type: InferenceExperimentDataStorageConfig object

Description

The description of the inference experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

EndpointMetadata

The metadata of the endpoint on which the inference experiment ran.

Type: EndpointMetadata object

KmsKey

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt data on the storage volume attached to the ML compute instance that hosts the endpoint. For more information, see CreateInferenceExperiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

LastModifiedTime

The timestamp at which you last modified the inference experiment.

Type: Timestamp

ModelVariants

An array of `ModelVariantConfigSummary` objects. There is one for each variant in the inference experiment. Each `ModelVariantConfigSummary` object in the array describes the infrastructure configuration for deploying the corresponding variant.

Type: Array of ModelVariantConfigSummary objects

Name

The name of the inference experiment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

RoleArn

The ARN of the IAM role that Amazon SageMaker can assume to access model artifacts and container images, and manage Amazon SageMaker Inference endpoints for model deployment.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Schedule

The duration for which the inference experiment ran or will run.

Type: InferenceExperimentSchedule object

ShadowModeConfig

The configuration of `ShadowMode` inference experiment type, which shows the production variant that takes all the inference requests, and the shadow variant to which Amazon SageMaker replicates a percentage of the inference requests. For the shadow variant it also shows the percentage of requests that Amazon SageMaker replicates.

Type: ShadowModeConfig object

Status

The status of the inference experiment. The following are the possible statuses for an inference experiment:

- `Creating` - Amazon SageMaker is creating your experiment.
- `Created` - Amazon SageMaker has finished the creation of your experiment and will begin the experiment at the scheduled time.
- `Updating` - When you make changes to your experiment, your experiment shows as updating.
- `Starting` - Amazon SageMaker is beginning your experiment.
- `Running` - Your experiment is in progress.
- `Stopping` - Amazon SageMaker is stopping your experiment.
- `Completed` - Your experiment has completed.
- `Cancelled` - When you conclude your experiment early using the StopInferenceExperiment API, or if any operation fails with an unexpected error, it shows as cancelled.

Type: String

Valid Values: `Creating | Created | Updating | Running | Starting | Stopping | Completed | Cancelled`

StatusReason

The error message or client-specified `Reason` from the StopInferenceExperiment API, that explains the status of the inference experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Type

The type of the inference experiment.

Type: String

Valid Values: `ShadowMode`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeInferenceRecommendationsJob

Service: Amazon SageMaker Service

Provides the results of the Inference Recommender job. One or more recommendation jobs are returned.

Request Syntax

```
{
   "JobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobName

The name of the job. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: Yes

Response Syntax

```
{
   "CompletionTime": number,
   "CreationTime": number,
   "EndpointPerformances": [
      {
         "EndpointInfo": {
            "EndpointName": "string"
         },
         "Metrics": {
            "MaxInvocations": number,
```

```
            "ModelLatency": number
         }
      }
   ],
   "FailureReason": "string",
   "InferenceRecommendations": [
      {
         "EndpointConfiguration": {
            "EndpointName": "string",
            "InitialInstanceCount": number,
            "InstanceType": "string",
            "ServerlessConfig": {
               "MaxConcurrency": number,
               "MemorySizeInMB": number,
               "ProvisionedConcurrency": number
            },
            "VariantName": "string"
         },
         "InvocationEndTime": number,
         "InvocationStartTime": number,
         "Metrics": {
            "CostPerHour": number,
            "CostPerInference": number,
            "CpuUtilization": number,
            "MaxInvocations": number,
            "MemoryUtilization": number,
            "ModelLatency": number,
            "ModelSetupTime": number
         },
         "ModelConfiguration": {
            "CompilationJobName": "string",
            "EnvironmentParameters": [
               {
                  "Key": "string",
                  "Value": "string",
                  "ValueType": "string"
               }
            ],
            "InferenceSpecificationName": "string"
         },
         "RecommendationId": "string"
      }
   ],
   "InputConfig": {
```

```
      "ContainerConfig": {
         "DataInputConfig": "string",
         "Domain": "string",
         "Framework": "string",
         "FrameworkVersion": "string",
         "NearestModelName": "string",
         "PayloadConfig": {
            "SamplePayloadUrl": "string",
            "SupportedContentTypes": [ "string" ]
         },
         "SupportedEndpointType": "string",
         "SupportedInstanceTypes": [ "string" ],
         "SupportedResponseMIMETypes": [ "string" ],
         "Task": "string"
      },
      "EndpointConfigurations": [
         {
            "EnvironmentParameterRanges": {
               "CategoricalParameterRanges": [
                  {
                     "Name": "string",
                     "Value": [ "string" ]
                  }
               ]
            },
            "InferenceSpecificationName": "string",
            "InstanceType": "string",
            "ServerlessConfig": {
               "MaxConcurrency": number,
               "MemorySizeInMB": number,
               "ProvisionedConcurrency": number
            }
         }
      ],
      "Endpoints": [
         {
            "EndpointName": "string"
         }
      ],
      "JobDurationInSeconds": number,
      "ModelName": "string",
      "ModelPackageVersionArn": "string",
      "ResourceLimit": {
         "MaxNumberOfTests": number,
```

```
            "MaxParallelOfTests": number
         },
         "TrafficPattern": {
            "Phases": [
               {
                  "DurationInSeconds": number,
                  "InitialNumberOfUsers": number,
                  "SpawnRate": number
               }
            ],
            "Stairs": {
               "DurationInSeconds": number,
               "NumberOfSteps": number,
               "UsersPerStep": number
            },
            "TrafficType": "string"
         },
         "VolumeKmsKeyId": "string",
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      },
      "JobArn": "string",
      "JobDescription": "string",
      "JobName": "string",
      "JobType": "string",
      "LastModifiedTime": number,
      "RoleArn": "string",
      "Status": "string",
      "StoppingConditions": {
         "FlatInvocations": "string",
         "MaxInvocations": number,
         "ModelLatencyThresholds": [
            {
               "Percentile": "string",
               "ValueInMilliseconds": number
            }
         ]
      }
   }
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CompletionTime

A timestamp that shows when the job completed.

Type: Timestamp

CreationTime

A timestamp that shows when the job was created.

Type: Timestamp

EndpointPerformances

The performance results from running an Inference Recommender job on an existing endpoint.

Type: Array of EndpointPerformance objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

FailureReason

If the job fails, provides information why the job failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

InferenceRecommendations

The recommendations made by Inference Recommender.

Type: Array of InferenceRecommendation objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

InputConfig

Returns information about the versioned model package Amazon Resource Name (ARN), the traffic pattern, and endpoint configurations you provided when you initiated the job.

Type: RecommendationJobInputConfig object

JobArn

The Amazon Resource Name (ARN) of the job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:inference-recommendations-job/.*`

JobDescription

The job description that you provided when you initiated the job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

JobName

The name of the job. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

JobType

The job type that you provided when you initiated the job.

Type: String

Valid Values: `Default | Advanced`

LastModifiedTime

A timestamp that shows when the job was last modified.

Type: Timestamp

RoleArn

The Amazon Resource Name (ARN) of the AWS Identity and Access Management (IAM) role you provided when you initiated the job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Status

The status of the job.

Type: String

Valid Values: `PENDING | IN_PROGRESS | COMPLETED | FAILED | STOPPING | STOPPED | DELETING | DELETED`

StoppingConditions

The stopping conditions that you provided when you initiated the job.

Type: RecommendationJobStoppingConditions object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeJob

Service: Amazon SageMaker Service

Returns detailed information about a job, including its current status, secondary status, configuration, and timestamps. Use `SecondaryStatus` for granular progress tracking and `SecondaryStatusTransitions` to see the full history of status changes with timestamps.

The following operations are related to `DescribeJob`:

- `CreateJob`
- `ListJobs`
- `StopJob`
- `DeleteJob`

Request Syntax

```
{
   "JobCategory": "string",
   "JobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobCategory

The category of the job.

Type: String

Valid Values: `AgentRFT | AgentRFTEvaluation`

Required: Yes

JobName

The name of the job to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "EndTime": number,
   "FailureReason": "string",
   "JobArn": "string",
   "JobCategory": "string",
   "JobConfigDocument": "string",
   "JobConfigSchemaVersion": "string",
   "JobName": "string",
   "JobStatus": "string",
   "LastModifiedTime": number,
   "RoleArn": "string",
   "SecondaryStatus": "string",
   "SecondaryStatusTransitions": [
      {
         "EndTime": number,
         "StartTime": number,
         "Status": "string",
         "StatusMessage": "string"
      }
   ],
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The date and time that the job was created.

Type: Timestamp

EndTime

The date and time that the job ended.

Type: Timestamp

FailureReason

If the job failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

JobArn

The Amazon Resource Name (ARN) of the job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:(aws[a-z\-]*):sagemaker:[a-z0-9\-]+:[0-9]{12}:job/[a-zA-Z0-9]+/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

JobCategory

The category of the job.

Type: String

Valid Values: `AgentRFT | AgentRFTEvaluation`

JobConfigDocument

The JSON configuration document for the job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 262144.

JobConfigSchemaVersion

The schema version used for the job configuration document.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 16.

Pattern: `\d+\.\d+\.\d+`

JobName

The name of the job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

JobStatus

The current status of the job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting | DeleteFailed`

LastModifiedTime

The date and time that the job was last modified.

Type: Timestamp

RoleArn

The ARN of the IAM role associated with the job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

SecondaryStatus

The detailed secondary status of the job, providing more granular information about the job's progress. Secondary statuses may change between releases.

Type: String

Valid Values: `Starting | Downloading | Training | Uploading | Stopping | Stopped | MaxRuntimeExceeded | Interrupted | Failed | Completed | Restarting | Pending | Evaluating | Deleting | DeleteFailed`

SecondaryStatusTransitions

A list of secondary status transitions for the job, with timestamps and optional status messages.

Type: Array of JobSecondaryStatusTransition objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Tags

The tags associated with the job.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeJobSchemaVersion

Service: Amazon SageMaker Service

Returns the JSON schema for a specified job category and schema version. Use this schema to validate your `JobConfigDocument` before calling `CreateJob`. If you don't specify a schema version, the latest version is returned. The schema defines required fields, allowed values, and constraints for the job configuration.

The following operations are related to `DescribeJobSchemaVersion`:

- `ListJobSchemaVersions`
- `CreateJob`

Request Syntax

```
{
   "JobCategory": "string",
   "JobConfigSchemaVersion": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobCategory

The category of the job schema to describe.

Type: String

Valid Values: `AgentRFT | AgentRFTEvaluation`

Required: Yes

JobConfigSchemaVersion

The version of the schema to retrieve. If not specified, the latest version is returned.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 16.

Pattern: `\d+\.\d+\.\d+`

Required: No

Response Syntax

```
{
   "JobCategory": "string",
   "JobConfigSchema": "string",
   "JobConfigSchemaVersion": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobCategory

The category of the job schema.

Type: String

Valid Values: `AgentRFT | AgentRFTEvaluation`

JobConfigSchema

The JSON schema document that defines the structure of the job configuration.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 262144.

JobConfigSchemaVersion

The version of the schema.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 16.

Pattern: `\d+\.\d+\.\d+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeLabelingJob

Service: Amazon SageMaker Service

Gets information about a labeling job.

Request Syntax

```
{
   "LabelingJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

LabelingJobName

The name of the labeling job to return information for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "FailureReason": "string",
   "HumanTaskConfig": {
      "AnnotationConsolidationConfig": {
         "AnnotationConsolidationLambdaArn": "string"
      },
      "MaxConcurrentTaskCount": number,
      "NumberOfHumanWorkersPerDataObject": number,
      "PreHumanTaskLambdaArn": "string",
      "PublicWorkforceTaskPrice": {
         "AmountInUsd": {
```

```
            "Cents": number,
            "Dollars": number,
            "TenthFractionsOfACent": number
         }
      },
      "TaskAvailabilityLifetimeInSeconds": number,
      "TaskDescription": "string",
      "TaskKeywords": [ "string" ],
      "TaskTimeLimitInSeconds": number,
      "TaskTitle": "string",
      "UiConfig": {
         "HumanTaskUiArn": "string",
         "UiTemplateS3Uri": "string"
      },
      "WorkteamArn": "string"
   },
   "InputConfig": {
      "DataAttributes": {
         "ContentClassifiers": [ "string" ]
      },
      "DataSource": {
         "S3DataSource": {
            "ManifestS3Uri": "string"
         },
         "SnsDataSource": {
            "SnsTopicArn": "string"
         }
      }
   },
   "JobReferenceCode": "string",
   "LabelAttributeName": "string",
   "LabelCategoryConfigS3Uri": "string",
   "LabelCounters": {
      "FailedNonRetryableError": number,
      "HumanLabeled": number,
      "MachineLabeled": number,
      "TotalLabeled": number,
      "Unlabeled": number
   },
   "LabelingJobAlgorithmsConfig": {
      "InitialActiveLearningModelArn": "string",
      "LabelingJobAlgorithmSpecificationArn": "string",
      "LabelingJobResourceConfig": {
         "VolumeKmsKeyId": "string",
```

```
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      }
   },
   "LabelingJobArn": "string",
   "LabelingJobName": "string",
   "LabelingJobOutput": {
      "FinalActiveLearningModelArn": "string",
      "OutputDatasetS3Uri": "string"
   },
   "LabelingJobStatus": "string",
   "LastModifiedTime": number,
   "OutputConfig": {
      "KmsKeyId": "string",
      "S3OutputPath": "string",
      "SnsTopicArn": "string"
   },
   "RoleArn": "string",
   "StoppingConditions": {
      "MaxHumanLabeledObjectCount": number,
      "MaxPercentageOfInputDatasetLabeled": number
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The date and time that the labeling job was created.

Type: Timestamp

FailureReason

If the job failed, the reason that it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

HumanTaskConfig

Configuration information required for human workers to complete a labeling task.

Type: HumanTaskConfig object

InputConfig

Input configuration information for the labeling job, such as the Amazon S3 location of the data objects and the location of the manifest file that describes the data objects.

Type: LabelingJobInputConfig object

JobReferenceCode

A unique identifier for work done as part of a labeling job.

Type: String

Length Constraints: Minimum length of 1.

Pattern: `.+`

LabelAttributeName

The attribute used as the label in the output manifest file.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 127.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,126}`

LabelCategoryConfigS3Uri

The S3 location of the JSON file that defines the categories used to label data objects. Please note the following label-category limits:

- Semantic segmentation labeling jobs using automated labeling: 20 labels

- Box bounding labeling jobs (all): 10 labels

The file is a JSON structure in the following format:

```
{
"document-version": "2018-11-28"
"labels": [
{
"label": "label 1"
},
{
"label": "label 2"
},
...
{
"label": "label n"
}
]
}
```

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

LabelCounters

Provides a breakdown of the number of data objects labeled by humans, the number of objects labeled by machine, the number of objects than couldn't be labeled, and the total number of objects labeled.

Type: LabelCounters object

LabelingJobAlgorithmsConfig

Configuration information for automated data labeling.

Type: LabelingJobAlgorithmsConfig object

LabelingJobArn

The Amazon Resource Name (ARN) of the labeling job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:labeling-job/.*`

LabelingJobName

The name assigned to the labeling job when it was created.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

LabelingJobOutput

The location of the output produced by the labeling job.

Type: LabelingJobOutput object

LabelingJobStatus

The processing status of the labeling job.

Type: String

Valid Values: `Initializing | InProgress | Completed | Failed | Stopping | Stopped`

LastModifiedTime

The date and time that the labeling job was last updated.

Type: Timestamp

OutputConfig

The location of the job's output data and the AWS Key Management Service key ID for the key used to encrypt the output data, if any.

Type: LabelingJobOutputConfig object

RoleArn

The Amazon Resource Name (ARN) that SageMaker assumes to perform tasks on your behalf during data labeling.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

StoppingConditions

A set of conditions for stopping a labeling job. If any of the conditions are met, the job is automatically stopped.

Type: LabelingJobStoppingConditions object

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeLineageGroup

Service: Amazon SageMaker Service

Provides a list of properties for the requested lineage group. For more information, see Cross-Account Lineage Tracking in the *Amazon SageMaker Developer Guide*.

Request Syntax

```
{
   "LineageGroupName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

LineageGroupName

The name of the lineage group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
```

```
      "UserProfileName": "string"
   },
   "CreationTime": number,
   "Description": "string",
   "DisplayName": "string",
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "LineageGroupArn": "string",
   "LineageGroupName": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

The creation time of lineage group.

Type: Timestamp

Description

The description of the lineage group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

DisplayName

The display name of the lineage group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

The last modified time of the lineage group.

Type: Timestamp

LineageGroupArn

The Amazon Resource Name (ARN) of the lineage group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:lineage-group/.*`

LineageGroupName

The name of the lineage group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeMlflowApp

Service: Amazon SageMaker Service

Returns information about an MLflow App.

Request Syntax

```
{
   "Arn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Arn

The ARN of the MLflow App for which to get information.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-app/.*`

Required: Yes

Response Syntax

```
{
   "AccountDefaultStatus": "string",
   "Arn": "string",
   "ArtifactStoreUri": "string",
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
```

```
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "CreationTime": number,
   "DefaultDomainIdList": [ "string" ],
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "MaintenanceStatus": "string",
   "MlflowVersion": "string",
   "ModelRegistrationMode": "string",
   "Name": "string",
   "RoleArn": "string",
   "Status": "string",
   "WeeklyMaintenanceWindowStart": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AccountDefaultStatus

Indicates whether this MLflow app is the default for the entire account.

Type: String

Valid Values: `ENABLED | DISABLED`

Arn

The ARN of the MLflow App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-app/.*`

ArtifactStoreUri

The S3 URI of the general purpose bucket used as the MLflow App artifact store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

The timestamp when the MLflow App was created.

Type: Timestamp

DefaultDomainIdList

List of SageMaker Domain IDs for which this MLflow App is the default.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

The timestamp when the MLflow App was last modified.

Type: Timestamp

MaintenanceStatus

Current maintenance status of the MLflow App.

Type: String

Valid Values: `MaintenanceInProgress | MaintenanceComplete | MaintenanceFailed`

MlflowVersion

The MLflow version used.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 16.

Pattern: `[0-9]*.[0-9]*.[0-9]*`

ModelRegistrationMode

Whether automatic registration of new MLflow models to the SageMaker Model Registry is enabled.

Type: String

Valid Values: `AutoModelRegistrationEnabled | AutoModelRegistrationDisabled`

Name

The name of the MLflow App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

RoleArn

The Amazon Resource Name (ARN) for an IAM role in your account that the MLflow App uses to access the artifact store in Amazon S3.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Status

The current creation status of the described MLflow App.

Type: String

Valid Values: `Creating | Created | CreateFailed | Updating | Updated | UpdateFailed | Deleting | DeleteFailed | Deleted`

WeeklyMaintenanceWindowStart

The day and time of the week when weekly maintenance occurs.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 9.

Pattern: `(Mon|Tue|Wed|Thu|Fri|Sat|Sun):([01]\d|2[0-3]):([0-5]\d)`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeMlflowTrackingServer

Service: Amazon SageMaker Service

Returns information about an MLflow Tracking Server.

Request Syntax

```
{
   "TrackingServerName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrackingServerName

The name of the MLflow Tracking Server to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: Yes

Response Syntax

```
{
   "ArtifactStoreUri": "string",
   "AutomaticModelRegistration": boolean,
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
```

```
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "CreationTime": number,
   "IsActive": "string",
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "MlflowVersion": "string",
   "RoleArn": "string",
   "S3BucketOwnerAccountId": "string",
   "S3BucketOwnerVerification": boolean,
   "TrackingServerArn": "string",
   "TrackingServerMaintenanceStatus": "string",
   "TrackingServerName": "string",
   "TrackingServerSize": "string",
   "TrackingServerStatus": "string",
   "TrackingServerUrl": "string",
   "WeeklyMaintenanceWindowStart": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ArtifactStoreUri

The S3 URI of the general purpose bucket used as the MLflow Tracking Server artifact store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

AutomaticModelRegistration

Whether automatic registration of new MLflow models to the SageMaker Model Registry is enabled.

Type: Boolean

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

The timestamp of when the described MLflow Tracking Server was created.

Type: Timestamp

IsActive

Whether the described MLflow Tracking Server is currently active.

Type: String

Valid Values: `Active | Inactive`

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

The timestamp of when the described MLflow Tracking Server was last modified.

Type: Timestamp

MlflowVersion

The MLflow version used for the described tracking server.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 16.

Pattern: `[0-9]*.[0-9]*.[0-9]*`

RoleArn

The Amazon Resource Name (ARN) for an IAM role in your account that the described MLflow Tracking Server uses to access the artifact store in Amazon S3.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

S3BucketOwnerAccountId

Expected AWS account ID that owns the Amazon S3 bucket for artifact storage.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `\d+`

S3BucketOwnerVerification

Whether Amazon S3 Bucket Ownership checks are enabled whenever the tracking server interacts with Amazon Amazon S3.

Type: Boolean

TrackingServerArn

The ARN of the described tracking server.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-tracking-server/.*`

TrackingServerMaintenanceStatus

The current maintenance status of the described MLflow Tracking Server.

Type: String

Valid Values: `MaintenanceInProgress | MaintenanceComplete | MaintenanceFailed`

TrackingServerName

The name of the described tracking server.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

TrackingServerSize

The size of the described tracking server.

Type: String

Valid Values: `Small | Medium | Large`

TrackingServerStatus

The current creation status of the described MLflow Tracking Server.

Type: String

Valid Values: `Creating | Created | CreateFailed | Updating | Updated | UpdateFailed | Deleting | DeleteFailed | Stopping | Stopped | StopFailed | Starting | Started | StartFailed | MaintenanceInProgress | MaintenanceComplete | MaintenanceFailed`

TrackingServerUrl

The URL to connect to the MLflow user interface for the described tracking server.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

WeeklyMaintenanceWindowStart

The day and time of the week when weekly maintenance occurs on the described tracking server.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 9.

Pattern: `(Mon|Tue|Wed|Thu|Fri|Sat|Sun):([01]\d|2[0-3]):([0-5]\d)`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeModel

Service: Amazon SageMaker Service

Describes a model that you created using the `CreateModel` API.

Request Syntax

```
{
   "ModelName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelName

The name of the model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

Response Syntax

```
{
   "Containers": [
      {
         "AdditionalModelDataSources": [
            {
               "ChannelName": "string",
               "S3DataSource": {
                  "CompressionType": "string",
                  "ETag": "string",
                  "HubAccessConfig": {
                     "HubContentArn": "string"
                  },
```

```
                "ManifestEtag": "string",
                "ManifestS3Uri": "string",
                "ModelAccessConfig": {
                   "AcceptEula": boolean
                },
                "S3DataType": "string",
                "S3Uri": "string"
             }
          }
       ],
       "ContainerHostname": "string",
       "ContainerMetricsConfig": {
          "MetricsEndpoints": [
             {
                "MetricPublishFrequencyInSeconds": number,
                "MetricsEndpointPath": "string"
             }
          ]
       },
       "Environment": {
          "string" : "string"
       },
       "Image": "string",
       "ImageConfig": {
          "RepositoryAccessMode": "string",
          "RepositoryAuthConfig": {
             "RepositoryCredentialsProviderArn": "string"
          }
       },
       "InferenceSpecificationName": "string",
       "Mode": "string",
       "ModelDataSource": {
          "S3DataSource": {
             "CompressionType": "string",
             "ETag": "string",
             "HubAccessConfig": {
                "HubContentArn": "string"
             },
             "ManifestEtag": "string",
             "ManifestS3Uri": "string",
             "ModelAccessConfig": {
                "AcceptEula": boolean
             },
             "S3DataType": "string",
```

```
            "S3Uri": "string"
         }
      },
      "ModelDataUrl": "string",
      "ModelPackageName": "string",
      "MultiModelConfig": {
         "ModelCacheSetting": "string"
      }
   }
],
"CreationTime": number,
"DeploymentRecommendation": {
   "RealTimeInferenceRecommendations": [
      {
         "Environment": {
            "string" : "string"
         },
         "InstanceType": "string",
         "RecommendationId": "string"
      }
   ],
   "RecommendationStatus": "string"
},
"EnableNetworkIsolation": boolean,
"ExecutionRoleArn": "string",
"InferenceExecutionConfig": {
   "Mode": "string"
},
"ModelArn": "string",
"ModelName": "string",
"PrimaryContainer": {
   "AdditionalModelDataSources": [
      {
         "ChannelName": "string",
         "S3DataSource": {
            "CompressionType": "string",
            "ETag": "string",
            "HubAccessConfig": {
               "HubContentArn": "string"
            },
            "ManifestEtag": "string",
            "ManifestS3Uri": "string",
            "ModelAccessConfig": {
               "AcceptEula": boolean
```

```
            },
            "S3DataType": "string",
            "S3Uri": "string"
         }
      }
   ],
   "ContainerHostname": "string",
   "ContainerMetricsConfig": {
      "MetricsEndpoints": [
         {
            "MetricPublishFrequencyInSeconds": number,
            "MetricsEndpointPath": "string"
         }
      ]
   },
   "Environment": {
      "string" : "string"
   },
   "Image": "string",
   "ImageConfig": {
      "RepositoryAccessMode": "string",
      "RepositoryAuthConfig": {
         "RepositoryCredentialsProviderArn": "string"
      }
   },
   "InferenceSpecificationName": "string",
   "Mode": "string",
   "ModelDataSource": {
      "S3DataSource": {
         "CompressionType": "string",
         "ETag": "string",
         "HubAccessConfig": {
            "HubContentArn": "string"
         },
         "ManifestEtag": "string",
         "ManifestS3Uri": "string",
         "ModelAccessConfig": {
            "AcceptEula": boolean
         },
         "S3DataType": "string",
         "S3Uri": "string"
      }
   },
   "ModelDataUrl": "string",
```

```
      "ModelPackageName": "string",
      "MultiModelConfig": {
         "ModelCacheSetting": "string"
      }
   },
   "VpcConfig": {
      "SecurityGroupIds": [ "string" ],
      "Subnets": [ "string" ]
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Containers

The containers in the inference pipeline.

Type: Array of ContainerDefinition objects

Array Members: Minimum number of 0 items. Maximum number of 15 items.

CreationTime

A timestamp that shows when the model was created.

Type: Timestamp

DeploymentRecommendation

A set of recommended deployment configurations for the model.

Type: DeploymentRecommendation object

EnableNetworkIsolation

If `True`, no inbound or outbound network calls can be made to or from the model container.

Type: Boolean

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you specified for the model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

InferenceExecutionConfig

Specifies details of how containers in a multi-container endpoint are called.

Type: InferenceExecutionConfig object

ModelArn

The Amazon Resource Name (ARN) of the model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:model/.*`

ModelName

Name of the SageMaker model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

PrimaryContainer

The location of the primary inference code, associated artifacts, and custom environment map that the inference code uses when it is deployed in production.

Type: ContainerDefinition object

VpcConfig

A VpcConfig object that specifies the VPC that this model has access to. For more information, see Protect Endpoints by Using an Amazon Virtual Private Cloud

Type: VpcConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeModelBiasJobDefinition

Service: Amazon SageMaker Service

Returns a description of a model bias job definition.

Request Syntax

```
{
   "JobDefinitionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobDefinitionName

The name of the model bias job definition. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "JobDefinitionArn": "string",
   "JobDefinitionName": "string",
   "JobResources": {
      "ClusterConfig": {
         "InstanceCount": number,
         "InstanceType": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
```

```
      }
   },
   "ModelBiasAppSpecification": {
      "ConfigUri": "string",
      "Environment": {
         "string" : "string"
      },
      "ImageUri": "string"
   },
   "ModelBiasBaselineConfig": {
      "BaseliningJobName": "string",
      "ConstraintsResource": {
         "S3Uri": "string"
      }
   },
   "ModelBiasJobInput": {
      "BatchTransformInput": {
         "DataCapturedDestinationS3Uri": "string",
         "DatasetFormat": {
            "Csv": {
               "Header": boolean
            },
            "Json": {
               "Line": boolean
            },
            "Parquet": {
            }
         },
         "EndTimeOffset": "string",
         "ExcludeFeaturesAttribute": "string",
         "FeaturesAttribute": "string",
         "InferenceAttribute": "string",
         "LocalPath": "string",
         "ProbabilityAttribute": "string",
         "ProbabilityThresholdAttribute": number,
         "S3DataDistributionType": "string",
         "S3InputMode": "string",
         "StartTimeOffset": "string"
      },
      "EndpointInput": {
         "EndpointName": "string",
         "EndTimeOffset": "string",
         "ExcludeFeaturesAttribute": "string",
         "FeaturesAttribute": "string",
```

```
            "InferenceAttribute": "string",
            "LocalPath": "string",
            "ProbabilityAttribute": "string",
            "ProbabilityThresholdAttribute": number,
            "S3DataDistributionType": "string",
            "S3InputMode": "string",
            "StartTimeOffset": "string"
         },
         "GroundTruthS3Input": {
            "S3Uri": "string"
         }
      },
      "ModelBiasJobOutputConfig": {
         "KmsKeyId": "string",
         "MonitoringOutputs": [
            {
               "S3Output": {
                  "LocalPath": "string",
                  "S3UploadMode": "string",
                  "S3Uri": "string"
               }
            }
         ]
      },
      "NetworkConfig": {
         "EnableInterContainerTrafficEncryption": boolean,
         "EnableNetworkIsolation": boolean,
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      },
      "RoleArn": "string",
      "StoppingCondition": {
         "MaxRuntimeInSeconds": number
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The time at which the model bias job was created.

Type: Timestamp

JobDefinitionArn

The Amazon Resource Name (ARN) of the model bias job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

JobDefinitionName

The name of the bias job definition. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

JobResources

Identifies the resources to deploy for a monitoring job.

Type: MonitoringResources object

ModelBiasAppSpecification

Configures the model bias job to run a specified Docker container image.

Type: ModelBiasAppSpecification object

ModelBiasBaselineConfig

The baseline configuration for a model bias job.

Type: ModelBiasBaselineConfig object

ModelBiasJobInput

Inputs for the model bias job.

Type: ModelBiasJobInput object

ModelBiasJobOutputConfig

The output configuration for monitoring jobs.

Type: MonitoringOutputConfig object

NetworkConfig

Networking options for a model bias job.

Type: MonitoringNetworkConfig object

RoleArn

The Amazon Resource Name (ARN) of the IAM role that has read permission to the input data location and write permission to the output data location in Amazon S3.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

StoppingCondition

A time limit for how long the monitoring job is allowed to run before stopping.

Type: MonitoringStoppingCondition object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeModelCard

Service: Amazon SageMaker Service

Describes the content, creation time, and security configuration of an Amazon SageMaker Model Card.

> **Important**
>
> To retrieve only metadata about a model card without requiring `kms:Decrypt` permission on the associated customer-managed AWS KMS key, set `IncludedData` to `MetadataOnly`. The default is `AllData`, which returns the full model card `Content` and requires `kms:Decrypt` permission when a customer-managed key is configured.

Request Syntax

```
{
   "IncludedData": "string",
   "ModelCardName": "string",
   "ModelCardVersion": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

IncludedData

Specifies the level of model card data to include in the response. Use this parameter to call `DescribeModelCard` without requiring `kms:Decrypt` permission on the customer-managed AWS KMS key.

- `AllData`: Returns the full model card `Content`. This option requires `kms:Decrypt` permission on the customer-managed key, if one is associated with the model card. This is the default.
- `MetadataOnly`: Returns the model card with sanitized `Content` that includes only a small set of unencrypted metadata fields. This option does not require `kms:Decrypt` permission. For the list of fields preserved in the response, see `Content`.

If you don't specify a value, SageMaker returns `AllData`.

Type: String

Valid Values: `AllData | MetadataOnly`

Required: No

ModelCardName

The name or Amazon Resource Name (ARN) of the model card to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:model-card/.*)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

ModelCardVersion

The version of the model card to describe. If a version is not provided, then the latest version of the model card is described.

Type: Integer

Required: No

Response Syntax

```
{
   "Content": "string",
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
```

```
   },
   "CreationTime": number,
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "ModelCardArn": "string",
   "ModelCardName": "string",
   "ModelCardProcessingStatus": "string",
   "ModelCardStatus": "string",
   "ModelCardVersion": number,
   "SecurityConfig": {
      "KmsKeyId": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Content

The content of the model card. Content is provided as a string in the model card JSON schema.

When you set `IncludedData` to `MetadataOnly` in the request, SageMaker returns a sanitized version of `Content` that includes only the following JSON paths, when present in the model card:

- `model_overview.model_id`
- `model_overview.model_name`
- `intended_uses.risk_rating`
- `model_package_details.model_package_group_name`
- `model_package_details.model_package_arn`

All other fields are removed from `Content` when `IncludedData` is `MetadataOnly`, including model description, training details, evaluation details, business details, and additional information. To retrieve the complete `Content`, set `IncludedData` to `AllData` or omit the parameter.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 100000.

Pattern: `.*`

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

The date and time the model card was created.

Type: Timestamp

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

The date and time the model card was last modified.

Type: Timestamp

ModelCardArn

The Amazon Resource Name (ARN) of the model card.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-card/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ModelCardName

The name of the model card.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ModelCardProcessingStatus

The processing status of model card deletion. The `ModelCardProcessingStatus` updates throughout the different deletion steps.

- `DeletePending`: Model card deletion request received.
- `DeleteInProgress`: Model card deletion is in progress.
- `ContentDeleted`: Deleted model card content.
- `ExportJobsDeleted`: Deleted all export jobs associated with the model card.
- `DeleteCompleted`: Successfully deleted the model card.
- `DeleteFailed`: The model card failed to delete.

Type: String

Valid Values: `DeleteInProgress | DeletePending | ContentDeleted | ExportJobsDeleted | DeleteCompleted | DeleteFailed`

ModelCardStatus

The approval status of the model card within your organization. Different organizations might have different criteria for model card review and approval.

- `Draft`: The model card is a work in progress.
- `PendingReview`: The model card is pending review.
- `Approved`: The model card is approved.
- `Archived`: The model card is archived. No more updates should be made to the model card, but it can still be exported.

Type: String

Valid Values: `Draft | PendingReview | Approved | Archived`

ModelCardVersion

The version of the model card.

Type: Integer

SecurityConfig

The security configuration used to protect model card content.

Type: ModelCardSecurityConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeModelCardExportJob

Service: Amazon SageMaker Service

Describes an Amazon SageMaker Model Card export job.

Request Syntax

```
{
   "ModelCardExportJobArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelCardExportJobArn

The Amazon Resource Name (ARN) of the model card export job to describe.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-card/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}/export-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreatedAt": number,
   "ExportArtifacts": {
      "S3ExportArtifacts": "string"
   },
   "FailureReason": "string",
   "LastModifiedAt": number,
   "ModelCardExportJobArn": "string",
```

```
   "ModelCardExportJobName": "string",
   "ModelCardName": "string",
   "ModelCardVersion": number,
   "OutputConfig": {
      "S3OutputPath": "string"
   },
   "Status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreatedAt

The date and time that the model export job was created.

Type: Timestamp

ExportArtifacts

The exported model card artifacts.

Type: ModelCardExportArtifacts object

FailureReason

The failure reason if the model export job fails.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

LastModifiedAt

The date and time that the model export job was last modified.

Type: Timestamp

ModelCardExportJobArn

The Amazon Resource Name (ARN) of the model card export job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-card/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}/export-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ModelCardExportJobName

The name of the model card export job to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ModelCardName

The name or Amazon Resource Name (ARN) of the model card that the model export job exports.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ModelCardVersion

The version of the model card that the model export job exports.

Type: Integer

OutputConfig

The export output details for the model card.

Type: ModelCardExportOutputConfig object

Status

The completion status of the model card export job.

- `InProgress`: The model card export job is in progress.
- `Completed`: The model card export job is complete.

- `Failed`: The model card export job failed. To see the reason for the failure, see the `FailureReason` field in the response to a `DescribeModelCardExportJob` call.

Type: String

Valid Values: `InProgress | Completed | Failed`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeModelExplainabilityJobDefinition

Service: Amazon SageMaker Service

Returns a description of a model explainability job definition.

Request Syntax

```
{
   "JobDefinitionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobDefinitionName

The name of the model explainability job definition. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "JobDefinitionArn": "string",
   "JobDefinitionName": "string",
   "JobResources": {
      "ClusterConfig": {
         "InstanceCount": number,
         "InstanceType": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
```

```
      }
   },
   "ModelExplainabilityAppSpecification": {
      "ConfigUri": "string",
      "Environment": {
         "string" : "string"
      },
      "ImageUri": "string"
   },
   "ModelExplainabilityBaselineConfig": {
      "BaseliningJobName": "string",
      "ConstraintsResource": {
         "S3Uri": "string"
      }
   },
   "ModelExplainabilityJobInput": {
      "BatchTransformInput": {
         "DataCapturedDestinationS3Uri": "string",
         "DatasetFormat": {
            "Csv": {
               "Header": boolean
            },
            "Json": {
               "Line": boolean
            },
            "Parquet": {
            }
         },
         "EndTimeOffset": "string",
         "ExcludeFeaturesAttribute": "string",
         "FeaturesAttribute": "string",
         "InferenceAttribute": "string",
         "LocalPath": "string",
         "ProbabilityAttribute": "string",
         "ProbabilityThresholdAttribute": number,
         "S3DataDistributionType": "string",
         "S3InputMode": "string",
         "StartTimeOffset": "string"
      },
      "EndpointInput": {
         "EndpointName": "string",
         "EndTimeOffset": "string",
         "ExcludeFeaturesAttribute": "string",
         "FeaturesAttribute": "string",
```

```
            "InferenceAttribute": "string",
            "LocalPath": "string",
            "ProbabilityAttribute": "string",
            "ProbabilityThresholdAttribute": number,
            "S3DataDistributionType": "string",
            "S3InputMode": "string",
            "StartTimeOffset": "string"
         }
      },
      "ModelExplainabilityJobOutputConfig": {
         "KmsKeyId": "string",
         "MonitoringOutputs": [
            {
               "S3Output": {
                  "LocalPath": "string",
                  "S3UploadMode": "string",
                  "S3Uri": "string"
               }
            }
         ]
      },
      "NetworkConfig": {
         "EnableInterContainerTrafficEncryption": boolean,
         "EnableNetworkIsolation": boolean,
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      },
      "RoleArn": "string",
      "StoppingCondition": {
         "MaxRuntimeInSeconds": number
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The time at which the model explainability job was created.

Type: Timestamp

JobDefinitionArn

The Amazon Resource Name (ARN) of the model explainability job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

JobDefinitionName

The name of the explainability job definition. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

JobResources

Identifies the resources to deploy for a monitoring job.

Type: MonitoringResources object

ModelExplainabilityAppSpecification

Configures the model explainability job to run a specified Docker container image.

Type: ModelExplainabilityAppSpecification object

ModelExplainabilityBaselineConfig

The baseline configuration for a model explainability job.

Type: ModelExplainabilityBaselineConfig object

ModelExplainabilityJobInput

Inputs for the model explainability job.

Type: ModelExplainabilityJobInput object

ModelExplainabilityJobOutputConfig

The output configuration for monitoring jobs.

Type: MonitoringOutputConfig object

NetworkConfig

Networking options for a model explainability job.

Type: MonitoringNetworkConfig object

RoleArn

The Amazon Resource Name (ARN) of the IAM role that has read permission to the input data location and write permission to the output data location in Amazon S3.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

StoppingCondition

A time limit for how long the monitoring job is allowed to run before stopping.

Type: MonitoringStoppingCondition object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeModelPackage

Service: Amazon SageMaker Service

Returns a description of the specified model package, which is used to create SageMaker models or list them on AWS Marketplace.

> **Important**
>
> If you provided a AWS KMS Key ID when you created your model package, you will see the KMS Decrypt API call in your CloudTrail logs when you use this API. To call this operation without requiring `kms:Decrypt` permission on the customer-managed key, set `IncludedData` to `MetadataOnly`; the response is returned with the embedded `ModelCard.ModelCardContent` field sanitized.

To create models in SageMaker, buyers can subscribe to model packages listed on AWS Marketplace.

Request Syntax

```
{
   "IncludedData": "string",
   "ModelPackageName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

IncludedData

Specifies the level of model package data to include in the response. Use this parameter to call `DescribeModelPackage` on a model package that has an associated model card without requiring `kms:Decrypt` permission on the customer-managed AWS KMS key associated with the embedded model card.

- `AllData`: Returns the full model package response, including the unredacted `ModelCard.ModelCardContent`. This option requires `kms:Decrypt` permission on the

customer-managed key, if one is associated with the embedded model card. This is the default.

- `MetadataOnly`: Returns the full model package response, but with the embedded `ModelCard.ModelCardContent` sanitized to include only a small set of unencrypted metadata fields. This option does not require `kms:Decrypt` permission. All other top-level response fields, including `InferenceSpecification`, `ModelMetrics`, `DriftCheckBaselines`, and `SecurityConfig`, are returned unchanged. For the list of fields preserved within `ModelCardContent`, see ModelCard.

If you don't specify a value, SageMaker returns `AllData`.

Type: String

Valid Values: `AllData | MetadataOnly`

Required: No

ModelPackageName

The name or Amazon Resource Name (ARN) of the model package to describe.

When you specify a name, the name must have 1 to 63 characters. Valid characters are a-z, A-Z, 0-9, and - (hyphen).

Type: String

Length Constraints: Minimum length of 1. Maximum length of 176.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*\/)?([a-zA-Z0-9]([a-zA-Z0-9-]){0,62})(?<!-)(\/[0-9]{1,9})?`

Required: Yes

Response Syntax

```
{
   "AdditionalInferenceSpecifications": [
      {
         "Containers": [
            {
               "AdditionalModelDataSources": [
                  {
                     "ChannelName": "string",
```

```
            "S3DataSource": {
               "CompressionType": "string",
               "ETag": "string",
               "HubAccessConfig": {
                  "HubContentArn": "string"
               },
               "ManifestEtag": "string",
               "ManifestS3Uri": "string",
               "ModelAccessConfig": {
                  "AcceptEula": boolean
               },
               "S3DataType": "string",
               "S3Uri": "string"
            }
         }
      ],
      "AdditionalS3DataSource": {
         "CompressionType": "string",
         "ETag": "string",
         "S3DataType": "string",
         "S3Uri": "string"
      },
      "BaseModel": {
         "HubContentName": "string",
         "HubContentVersion": "string",
         "RecipeName": "string"
      },
      "ContainerHostname": "string",
      "Environment": {
         "string" : "string"
      },
      "Framework": "string",
      "FrameworkVersion": "string",
      "Image": "string",
      "ImageDigest": "string",
      "IsCheckpoint": boolean,
      "ModelDataETag": "string",
      "ModelDataSource": {
         "S3DataSource": {
            "CompressionType": "string",
            "ETag": "string",
            "HubAccessConfig": {
               "HubContentArn": "string"
            },
```

```
                    "ManifestEtag": "string",
                    "ManifestS3Uri": "string",
                    "ModelAccessConfig": {
                       "AcceptEula": boolean
                    },
                    "S3DataType": "string",
                    "S3Uri": "string"
                 }
              },
              "ModelDataUrl": "string",
              "ModelInput": {
                 "DataInputConfig": "string"
              },
              "NearestModelName": "string",
              "ProductId": "string"
           }
        ],
        "Description": "string",
        "Name": "string",
        "SupportedContentTypes": [ "string" ],
        "SupportedRealtimeInferenceInstanceTypes": [ "string" ],
        "SupportedResponseMIMETypes": [ "string" ],
        "SupportedTransformInstanceTypes": [ "string" ]
     }
  ],
  "ApprovalDescription": "string",
  "CertifyForMarketplace": boolean,
  "CreatedBy": {
     "DomainId": "string",
     "IamIdentity": {
        "Arn": "string",
        "PrincipalId": "string",
        "SourceIdentity": "string"
     },
     "UserProfileArn": "string",
     "UserProfileName": "string"
  },
  "CreationTime": number,
  "CustomerMetadataProperties": {
     "string" : "string"
  },
  "Domain": "string",
  "DriftCheckBaselines": {
     "Bias": {
```

```
         "ConfigFile": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         },
         "PostTrainingConstraints": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         },
         "PreTrainingConstraints": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
      },
      "Explainability": {
         "ConfigFile": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         },
         "Constraints": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
      },
      "ModelDataQuality": {
         "Constraints": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         },
         "Statistics": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
      },
      "ModelQuality": {
         "Constraints": {
            "ContentDigest": "string",
            "ContentType": "string",
```

```
            "S3Uri": "string"
         },
         "Statistics": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
      }
   },
   "InferenceSpecification": {
      "Containers": [
         {
            "AdditionalModelDataSources": [
               {
                  "ChannelName": "string",
                  "S3DataSource": {
                     "CompressionType": "string",
                     "ETag": "string",
                     "HubAccessConfig": {
                        "HubContentArn": "string"
                     },
                     "ManifestEtag": "string",
                     "ManifestS3Uri": "string",
                     "ModelAccessConfig": {
                        "AcceptEula": boolean
                     },
                     "S3DataType": "string",
                     "S3Uri": "string"
                  }
               }
            ],
            "AdditionalS3DataSource": {
               "CompressionType": "string",
               "ETag": "string",
               "S3DataType": "string",
               "S3Uri": "string"
            },
            "BaseModel": {
               "HubContentName": "string",
               "HubContentVersion": "string",
               "RecipeName": "string"
            },
            "ContainerHostname": "string",
            "Environment": {
```

```
            "string" : "string"
         },
         "Framework": "string",
         "FrameworkVersion": "string",
         "Image": "string",
         "ImageDigest": "string",
         "IsCheckpoint": boolean,
         "ModelDataETag": "string",
         "ModelDataSource": {
            "S3DataSource": {
               "CompressionType": "string",
               "ETag": "string",
               "HubAccessConfig": {
                  "HubContentArn": "string"
               },
               "ManifestEtag": "string",
               "ManifestS3Uri": "string",
               "ModelAccessConfig": {
                  "AcceptEula": boolean
               },
               "S3DataType": "string",
               "S3Uri": "string"
            }
         },
         "ModelDataUrl": "string",
         "ModelInput": {
            "DataInputConfig": "string"
         },
         "NearestModelName": "string",
         "ProductId": "string"
      }
   ],
   "SupportedContentTypes": [ "string" ],
   "SupportedRealtimeInferenceInstanceTypes": [ "string" ],
   "SupportedResponseMIMETypes": [ "string" ],
   "SupportedTransformInstanceTypes": [ "string" ]
},
"LastModifiedBy": {
   "DomainId": "string",
   "IamIdentity": {
      "Arn": "string",
      "PrincipalId": "string",
      "SourceIdentity": "string"
   },
```

```
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "ManagedStorageType": "string",
   "MetadataProperties": {
      "CommitId": "string",
      "GeneratedBy": "string",
      "ProjectId": "string",
      "Repository": "string"
   },
   "ModelApprovalStatus": "string",
   "ModelCard": {
      "ModelCardContent": "string",
      "ModelCardStatus": "string"
   },
   "ModelLifeCycle": {
      "Stage": "string",
      "StageDescription": "string",
      "StageStatus": "string"
   },
   "ModelMetrics": {
      "Bias": {
         "PostTrainingReport": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         },
         "PreTrainingReport": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         },
         "Report": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
      },
      "Explainability": {
         "Report": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
```

```
            }
         },
         "ModelDataQuality": {
            "Constraints": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            },
            "Statistics": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            }
         },
         "ModelQuality": {
            "Constraints": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            },
            "Statistics": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            }
         }
      },
      "ModelPackageArn": "string",
      "ModelPackageDescription": "string",
      "ModelPackageGroupName": "string",
      "ModelPackageName": "string",
      "ModelPackageRegistrationType": "string",
      "ModelPackageStatus": "string",
      "ModelPackageStatusDetails": {
         "ImageScanStatuses": [
            {
               "FailureReason": "string",
               "Name": "string",
               "Status": "string"
            }
         ],
         "ValidationStatuses": [
            {
               "FailureReason": "string",
```

```
            "Name": "string",
            "Status": "string"
         }
      ]
   },
   "ModelPackageVersion": number,
   "SamplePayloadUrl": "string",
   "SecurityConfig": {
      "KmsKeyId": "string"
   },
   "SkipModelValidation": "string",
   "SourceAlgorithmSpecification": {
      "SourceAlgorithms": [
         {
            "AlgorithmName": "string",
            "ModelDataETag": "string",
            "ModelDataSource": {
               "S3DataSource": {
                  "CompressionType": "string",
                  "ETag": "string",
                  "HubAccessConfig": {
                     "HubContentArn": "string"
                  },
                  "ManifestEtag": "string",
                  "ManifestS3Uri": "string",
                  "ModelAccessConfig": {
                     "AcceptEula": boolean
                  },
                  "S3DataType": "string",
                  "S3Uri": "string"
               }
            },
            "ModelDataUrl": "string"
         }
      ]
   },
   "SourceUri": "string",
   "Task": "string",
   "ValidationSpecification": {
      "ValidationProfiles": [
         {
            "ProfileName": "string",
            "TransformJobDefinition": {
               "BatchStrategy": "string",
```

```
                "Environment": {
                   "string" : "string"
                },
                "MaxConcurrentTransforms": number,
                "MaxPayloadInMB": number,
                "TransformInput": {
                   "CompressionType": "string",
                   "ContentType": "string",
                   "DataSource": {
                      "S3DataSource": {
                         "S3DataType": "string",
                         "S3Uri": "string"
                      }
                   },
                   "SplitType": "string"
                },
                "TransformOutput": {
                   "Accept": "string",
                   "AssembleWith": "string",
                   "KmsKeyId": "string",
                   "S3OutputPath": "string"
                },
                "TransformResources": {
                   "InstanceCount": number,
                   "InstanceType": "string",
                   "TransformAmiVersion": "string",
                   "VolumeKmsKeyId": "string"
                }
             }
          }
       ],
       "ValidationRole": "string"
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AdditionalInferenceSpecifications

An array of additional Inference Specification objects. Each additional Inference Specification specifies artifacts based on this model package that can be used on inference endpoints. Generally used with SageMaker Neo to store the compiled artifacts.

Type: Array of AdditionalInferenceSpecificationDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 15 items.

ApprovalDescription

A description provided for the model approval.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

CertifyForMarketplace

Whether the model package is certified for listing on AWS Marketplace.

Type: Boolean

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

A timestamp specifying when the model package was created.

Type: Timestamp

CustomerMetadataProperties

The metadata properties associated with the model package versions.

Type: String to string map

Map Entries: Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `([\p{L}\p{Z}\p{N}_.:\/=+\-@]*)${1,128}`

Value Length Constraints: Minimum length of 1. Maximum length of 256.

Value Pattern: `([\p{L}\p{Z}\p{N}_.:\/=+\-@]*)${1,256}`

Domain

The machine learning domain of the model package you specified. Common machine learning domains include computer vision and natural language processing.

Type: String

DriftCheckBaselines

Represents the drift check baselines that can be used when the model monitor is set using the model package. For more information, see the topic on Drift Detection against Previous Baselines in SageMaker Pipelines in the *Amazon SageMaker Developer Guide*.

Type: DriftCheckBaselines object

InferenceSpecification

Details about inference jobs that you can run with models based on this model package.

Type: InferenceSpecification object

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

The last time that the model package was modified.

Type: Timestamp

ManagedStorageType

The storage type of the model package.

Type: String

Valid Values: `Restricted`

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

ModelApprovalStatus

The approval status of the model package.

Type: String

Valid Values: `Approved | Rejected | PendingManualApproval`

ModelCard

The model card associated with the model package. Since `ModelPackageModelCard` is tied to a model package, it is a specific usage of a model card and its schema is simplified compared to the schema of `ModelCard`. The `ModelPackageModelCard` schema does not include `model_package_details`, and `model_overview` is composed of the `model_creator` and `model_artifact` properties. For more information about the model package model card schema, see Model package model card schema. For more information about the model card associated with the model package, see View the Details of a Model Version.

When you set `IncludedData` to `MetadataOnly` in the request, `ModelCardStatus` is preserved and `ModelCardContent` is sanitized to include only the following JSON paths, when present in the model card:

- `model_overview.model_id`
- `model_overview.model_name`
- `intended_uses.risk_rating`
- `model_package_details.model_package_group_name`
- `model_package_details.model_package_arn`

Because the `ModelPackageModelCard` schema does not include `model_package_details` and limits `model_overview` to `model_creator` and `model_artifact`, the sanitized `ModelCardContent` for a model package typically contains only `intended_uses.risk_rating` if it was provided when the model card was created. To retrieve the complete `ModelCardContent`, set `IncludedData` to `AllData` or omit the parameter.

Type: ModelPackageModelCard object

ModelLifeCycle

A structure describing the current state of the model in its life cycle.

Type: ModelLifeCycle object

ModelMetrics

Metrics for the model.

Type: ModelMetrics object

ModelPackageArn

The Amazon Resource Name (ARN) of the model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

ModelPackageDescription

A brief summary of the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

ModelPackageGroupName

If the model is a versioned model, the name of the model group that the versioned model belongs to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ModelPackageName

The name of the model package being described.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ModelPackageRegistrationType

The package registration type of the model package output.

Type: String

Valid Values: `Logged | Registered`

ModelPackageStatus

The current status of the model package.

Type: String

Valid Values: `Pending | InProgress | Completed | Failed | Deleting`

ModelPackageStatusDetails

Details about the current status of the model package.

Type: ModelPackageStatusDetails object

ModelPackageVersion

The version of the model package.

Type: Integer

Valid Range: Minimum value of 1.

SamplePayloadUrl

The Amazon Simple Storage Service (Amazon S3) path where the sample payload are stored. This path points to a single gzip compressed tar archive (.tar.gz suffix).

Type: String

SecurityConfig

The AWS KMS Key ID (`KMSKeyId`) used for encryption of model package information.

Type: ModelPackageSecurityConfig object

SkipModelValidation

Indicates if you want to skip model validation.

Type: String

Valid Values: `All | None`

SourceAlgorithmSpecification

Details about the algorithm that was used to create the model package.

Type: SourceAlgorithmSpecification object

SourceUri

The URI of the source for the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{N}\p{P}]{0,1024}`

Task

The machine learning task you specified that your model package accomplishes. Common machine learning tasks include object detection and image classification.

Type: String

ValidationSpecification

Configurations for one or more transform jobs that SageMaker runs to test the model package.

Type: ModelPackageValidationSpecification object

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeModelPackageGroup

Service: Amazon SageMaker Service

Gets a description for the specified model group.

Request Syntax

```
{
   "ModelPackageGroupName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelPackageGroupName

The name of the model group to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 170.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*\/)?([a-zA-Z0-9]([a-zA-Z0-9-]){0,62})(?<!-)`

Required: Yes

Response Syntax

```
{
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
```

```
      "UserProfileName": "string"
   },
   "CreationTime": number,
   "ManagedConfiguration": {
      "ManagedStorageType": "string"
   },
   "ModelPackageGroupArn": "string",
   "ModelPackageGroupDescription": "string",
   "ModelPackageGroupName": "string",
   "ModelPackageGroupStatus": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

The time that the model group was created.

Type: Timestamp

ManagedConfiguration

The managed configuration of the model package group.

Type: ManagedConfiguration object

ModelPackageGroupArn

The Amazon Resource Name (ARN) of the model group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package-group/[\S]{1,2048}`

ModelPackageGroupDescription

A description of the model group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

ModelPackageGroupName

The name of the model group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ModelPackageGroupStatus

The status of the model group.

Type: String

Valid Values: `Pending | InProgress | Completed | Failed | Deleting | DeleteFailed`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeModelQualityJobDefinition

Service: Amazon SageMaker Service

Returns a description of a model quality job definition.

Request Syntax

```
{
   "JobDefinitionName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobDefinitionName

The name of the model quality job. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "JobDefinitionArn": "string",
   "JobDefinitionName": "string",
   "JobResources": {
      "ClusterConfig": {
         "InstanceCount": number,
         "InstanceType": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
```

```
      }
   },
   "ModelQualityAppSpecification": {
      "ContainerArguments": [ "string" ],
      "ContainerEntrypoint": [ "string" ],
      "Environment": {
         "string" : "string"
      },
      "ImageUri": "string",
      "PostAnalyticsProcessorSourceUri": "string",
      "ProblemType": "string",
      "RecordPreprocessorSourceUri": "string"
   },
   "ModelQualityBaselineConfig": {
      "BaseliningJobName": "string",
      "ConstraintsResource": {
         "S3Uri": "string"
      }
   },
   "ModelQualityJobInput": {
      "BatchTransformInput": {
         "DataCapturedDestinationS3Uri": "string",
         "DatasetFormat": {
            "Csv": {
               "Header": boolean
            },
            "Json": {
               "Line": boolean
            },
            "Parquet": {
            }
         },
         "EndTimeOffset": "string",
         "ExcludeFeaturesAttribute": "string",
         "FeaturesAttribute": "string",
         "InferenceAttribute": "string",
         "LocalPath": "string",
         "ProbabilityAttribute": "string",
         "ProbabilityThresholdAttribute": number,
         "S3DataDistributionType": "string",
         "S3InputMode": "string",
         "StartTimeOffset": "string"
      },
      "EndpointInput": {
```

```
            "EndpointName": "string",
            "EndTimeOffset": "string",
            "ExcludeFeaturesAttribute": "string",
            "FeaturesAttribute": "string",
            "InferenceAttribute": "string",
            "LocalPath": "string",
            "ProbabilityAttribute": "string",
            "ProbabilityThresholdAttribute": number,
            "S3DataDistributionType": "string",
            "S3InputMode": "string",
            "StartTimeOffset": "string"
         },
         "GroundTruthS3Input": {
            "S3Uri": "string"
         }
      },
      "ModelQualityJobOutputConfig": {
         "KmsKeyId": "string",
         "MonitoringOutputs": [
            {
               "S3Output": {
                  "LocalPath": "string",
                  "S3UploadMode": "string",
                  "S3Uri": "string"
               }
            }
         ]
      },
      "NetworkConfig": {
         "EnableInterContainerTrafficEncryption": boolean,
         "EnableNetworkIsolation": boolean,
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      },
      "RoleArn": "string",
      "StoppingCondition": {
         "MaxRuntimeInSeconds": number
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The time at which the model quality job was created.

Type: Timestamp

JobDefinitionArn

The Amazon Resource Name (ARN) of the model quality job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

JobDefinitionName

The name of the quality job definition. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

JobResources

Identifies the resources to deploy for a monitoring job.

Type: MonitoringResources object

ModelQualityAppSpecification

Configures the model quality job to run a specified Docker container image.

Type: ModelQualityAppSpecification object

ModelQualityBaselineConfig

The baseline configuration for a model quality job.

Type: ModelQualityBaselineConfig object

ModelQualityJobInput

Inputs for the model quality job.

Type: ModelQualityJobInput object

ModelQualityJobOutputConfig

The output configuration for monitoring jobs.

Type: MonitoringOutputConfig object

NetworkConfig

Networking options for a model quality job.

Type: MonitoringNetworkConfig object

RoleArn

The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker AI can assume to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

StoppingCondition

A time limit for how long the monitoring job is allowed to run before stopping.

Type: MonitoringStoppingCondition object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeMonitoringSchedule

Service: Amazon SageMaker Service

Describes the schedule for a monitoring job.

Request Syntax

```
{
   "MonitoringScheduleName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MonitoringScheduleName

Name of a previously created monitoring schedule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "EndpointName": "string",
   "FailureReason": "string",
   "LastModifiedTime": number,
   "LastMonitoringExecutionSummary": {
      "CreationTime": number,
      "EndpointName": "string",
      "FailureReason": "string",
      "LastModifiedTime": number,
      "MonitoringExecutionStatus": "string",
      "MonitoringJobDefinitionName": "string",
```

```
      "MonitoringScheduleName": "string",
      "MonitoringType": "string",
      "ProcessingJobArn": "string",
      "ScheduledTime": number
   },
   "MonitoringScheduleArn": "string",
   "MonitoringScheduleConfig": {
      "MonitoringJobDefinition": {
         "BaselineConfig": {
            "BaseliningJobName": "string",
            "ConstraintsResource": {
               "S3Uri": "string"
            },
            "StatisticsResource": {
               "S3Uri": "string"
            }
         },
         "Environment": {
            "string" : "string"
         },
         "MonitoringAppSpecification": {
            "ContainerArguments": [ "string" ],
            "ContainerEntrypoint": [ "string" ],
            "ImageUri": "string",
            "PostAnalyticsProcessorSourceUri": "string",
            "RecordPreprocessorSourceUri": "string"
         },
         "MonitoringInputs": [
            {
               "BatchTransformInput": {
                  "DataCapturedDestinationS3Uri": "string",
                  "DatasetFormat": {
                     "Csv": {
                        "Header": boolean
                     },
                     "Json": {
                        "Line": boolean
                     },
                     "Parquet": {
                     }
                  },
                  "EndTimeOffset": "string",
                  "ExcludeFeaturesAttribute": "string",
                  "FeaturesAttribute": "string",
```

```
                "InferenceAttribute": "string",
                "LocalPath": "string",
                "ProbabilityAttribute": "string",
                "ProbabilityThresholdAttribute": number,
                "S3DataDistributionType": "string",
                "S3InputMode": "string",
                "StartTimeOffset": "string"
            },
            "EndpointInput": {
                "EndpointName": "string",
                "EndTimeOffset": "string",
                "ExcludeFeaturesAttribute": "string",
                "FeaturesAttribute": "string",
                "InferenceAttribute": "string",
                "LocalPath": "string",
                "ProbabilityAttribute": "string",
                "ProbabilityThresholdAttribute": number,
                "S3DataDistributionType": "string",
                "S3InputMode": "string",
                "StartTimeOffset": "string"
            }
        }
    ],
    "MonitoringOutputConfig": {
        "KmsKeyId": "string",
        "MonitoringOutputs": [
            {
                "S3Output": {
                    "LocalPath": "string",
                    "S3UploadMode": "string",
                    "S3Uri": "string"
                }
            }
        ]
    },
    "MonitoringResources": {
        "ClusterConfig": {
            "InstanceCount": number,
            "InstanceType": "string",
            "VolumeKmsKeyId": "string",
            "VolumeSizeInGB": number
        }
    },
    "NetworkConfig": {
```

```
            "EnableInterContainerTrafficEncryption": boolean,
            "EnableNetworkIsolation": boolean,
            "VpcConfig": {
               "SecurityGroupIds": [ "string" ],
               "Subnets": [ "string" ]
            }
         },
         "RoleArn": "string",
         "StoppingCondition": {
            "MaxRuntimeInSeconds": number
         }
      },
      "MonitoringJobDefinitionName": "string",
      "MonitoringType": "string",
      "ScheduleConfig": {
         "DataAnalysisEndTime": "string",
         "DataAnalysisStartTime": "string",
         "ScheduleExpression": "string"
      }
   },
   "MonitoringScheduleName": "string",
   "MonitoringScheduleStatus": "string",
   "MonitoringType": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The time at which the monitoring job was created.

Type: Timestamp

EndpointName

The name of the endpoint for the monitoring job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

FailureReason

A string, up to one KB in size, that contains the reason a monitoring job failed, if it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

LastModifiedTime

The time at which the monitoring job was last modified.

Type: Timestamp

LastMonitoringExecutionSummary

Describes metadata on the last execution to run, if there was one.

Type: MonitoringExecutionSummary object

MonitoringScheduleArn

The Amazon Resource Name (ARN) of the monitoring schedule.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

MonitoringScheduleConfig

The configuration object that specifies the monitoring schedule and defines the monitoring job.

Type: MonitoringScheduleConfig object

MonitoringScheduleName

Name of the monitoring schedule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

MonitoringScheduleStatus

The status of an monitoring job.

Type: String

Valid Values: `Pending | Failed | Scheduled | Stopped`

MonitoringType

The type of the monitoring job that this schedule runs. This is one of the following values.

- `DATA_QUALITY` - The schedule is for a data quality monitoring job.
- `MODEL_QUALITY` - The schedule is for a model quality monitoring job.
- `MODEL_BIAS` - The schedule is for a bias monitoring job.
- `MODEL_EXPLAINABILITY` - The schedule is for an explainability monitoring job.

Type: String

Valid Values: `DataQuality | ModelQuality | ModelBias | ModelExplainability`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeNotebookInstance

Service: Amazon SageMaker Service

Returns information about a notebook instance.

Request Syntax

```
{
   "NotebookInstanceName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

NotebookInstanceName

The name of the notebook instance that you want information about.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

Response Syntax

```
{
   "AcceleratorTypes": [ "string" ],
   "AdditionalCodeRepositories": [ "string" ],
   "CreationTime": number,
   "DefaultCodeRepository": "string",
   "DirectInternetAccess": "string",
   "FailureReason": "string",
   "InstanceMetadataServiceConfiguration": {
      "MinimumInstanceMetadataServiceVersion": "string"
   },
```

```
    "InstanceType": "string",
    "IpAddressType": "string",
    "KmsKeyId": "string",
    "LastModifiedTime": number,
    "NetworkInterfaceId": "string",
    "NotebookInstanceArn": "string",
    "NotebookInstanceLifecycleConfigName": "string",
    "NotebookInstanceName": "string",
    "NotebookInstanceStatus": "string",
    "PlatformIdentifier": "string",
    "RoleArn": "string",
    "RootAccess": "string",
    "SecurityGroups": [ "string" ],
    "SubnetId": "string",
    "Url": "string",
    "VolumeSizeInGB": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AcceleratorTypes

This parameter is no longer supported. Elastic Inference (EI) is no longer available.

This parameter was used to specify a list of the EI instance types associated with this notebook instance.

Type: Array of strings

Valid Values: `ml.eia1.medium | ml.eia1.large | ml.eia1.xlarge | ml.eia2.medium | ml.eia2.large | ml.eia2.xlarge`

AdditionalCodeRepositories

An array of up to three Git repositories associated with the notebook instance. These can be either the names of Git repositories stored as resources in your account, or the URL of Git repositories in AWS CodeCommit or in any other Git repository. These repositories are cloned at the same level as the default repository of your notebook instance. For more information, see Associating Git Repositories with SageMaker AI Notebook Instances.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 3 items.

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `https://([^/]+)/?(.*)$|^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

CreationTime

A timestamp. Use this parameter to return the time when the notebook instance was created

Type: Timestamp

DefaultCodeRepository

The Git repository associated with the notebook instance as its default code repository. This can be either the name of a Git repository stored as a resource in your account, or the URL of a Git repository in AWS CodeCommit or in any other Git repository. When you open a notebook instance, it opens in the directory that contains this repository. For more information, see Associating Git Repositories with SageMaker AI Notebook Instances.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `https://([^/]+)/?(.*)$|^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

DirectInternetAccess

Describes whether SageMaker AI provides internet access to the notebook instance. If this value is set to *Disabled*, the notebook instance does not have internet access, and cannot connect to SageMaker AI training and endpoint services.

For more information, see Notebook Instances Are Internet-Enabled by Default.

Type: String

Valid Values: `Enabled | Disabled`

FailureReason

If status is `Failed`, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

InstanceMetadataServiceConfiguration

Information on the IMDS configuration of the notebook instance

Type: InstanceMetadataServiceConfiguration object

InstanceType

The type of ML compute instance running on the notebook instance.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.8xlarge | ml.m5d.12xlarge | ml.m5d.16xlarge | ml.m5d.24xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p3dn.24xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p6-b200.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c6i.large`

`| ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.m6id.large | ml.m6id.xlarge | ml.m6id.2xlarge | ml.m6id.4xlarge | ml.m6id.8xlarge | ml.m6id.12xlarge | ml.m6id.16xlarge | ml.m6id.24xlarge | ml.m6id.32xlarge | ml.c6id.large | ml.c6id.xlarge | ml.c6id.2xlarge | ml.c6id.4xlarge | ml.c6id.8xlarge | ml.c6id.12xlarge | ml.c6id.16xlarge | ml.c6id.24xlarge | ml.c6id.32xlarge | ml.r6id.large | ml.r6id.xlarge | ml.r6id.2xlarge | ml.r6id.4xlarge | ml.r6id.8xlarge | ml.r6id.12xlarge | ml.r6id.16xlarge | ml.r6id.24xlarge | ml.r6id.32xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.p5.4xlarge | ml.p5en.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge`

IpAddressType

The IP address type configured for the notebook instance. Returns `ipv4` for IPv4-only connectivity or `dualstack` for both IPv4 and IPv6 connectivity.

Type: String

Valid Values: `ipv4 | dualstack`

KmsKeyId

The AWS KMS key ID SageMaker AI uses to encrypt data when storing it on the ML storage volume attached to the instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

LastModifiedTime

A timestamp. Use this parameter to retrieve the time when the notebook instance was last modified.

Type: Timestamp

NetworkInterfaceId

The network interface IDs that SageMaker AI created at the time of creating the instance.

Type: String

NotebookInstanceArn

The Amazon Resource Name (ARN) of the notebook instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

NotebookInstanceLifecycleConfigName

Returns the name of a notebook instance lifecycle configuration.

For information about notebook instance lifestyle configurations, see Step 2.1: (Optional) Customize a Notebook Instance

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

NotebookInstanceName

The name of the SageMaker AI notebook instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

NotebookInstanceStatus

The status of the notebook instance.

Type: String

Valid Values: `Pending | InService | Stopping | Stopped | Failed | Deleting | Updating`

PlatformIdentifier

The platform identifier of the notebook instance runtime environment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 20.

Pattern: `(notebook-al1-v1|notebook-al2-v1|notebook-al2-v2|notebook-al2-v3|notebook-al2023-v1)`

RoleArn

The Amazon Resource Name (ARN) of the IAM role associated with the instance.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

RootAccess

Whether root access is enabled or disabled for users of the notebook instance.

> **Note**
>
> Lifecycle configurations need root access to be able to set up a notebook instance. Because of this, lifecycle configurations associated with a notebook instance always run with root access even if you disable root access for users.

Type: String

Valid Values: `Enabled | Disabled`

SecurityGroups

The IDs of the VPC security groups.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

SubnetId

The ID of the VPC subnet.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Url

The URL that you use to connect to the Jupyter notebook that is running in your notebook instance.

Type: String

VolumeSizeInGB

The size, in GB, of the ML storage volume attached to the notebook instance.

Type: Integer

Valid Range: Minimum value of 5. Maximum value of 16384.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeNotebookInstanceLifecycleConfig

Service: Amazon SageMaker Service

Returns a description of a notebook instance lifecycle configuration.

For information about notebook instance lifestyle configurations, see Step 2.1: (Optional) Customize a Notebook Instance.

Request Syntax

```
{
   "NotebookInstanceLifecycleConfigName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

NotebookInstanceLifecycleConfigName

The name of the lifecycle configuration to describe.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "LastModifiedTime": number,
   "NotebookInstanceLifecycleConfigArn": "string",
   "NotebookInstanceLifecycleConfigName": "string",
   "OnCreate": [
      {
```

```
            "Content": "string"
         }
      ],
      "OnStart": [
         {
            "Content": "string"
         }
      ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

A timestamp that tells when the lifecycle configuration was created.

Type: Timestamp

LastModifiedTime

A timestamp that tells when the lifecycle configuration was last modified.

Type: Timestamp

NotebookInstanceLifecycleConfigArn

The Amazon Resource Name (ARN) of the lifecycle configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

NotebookInstanceLifecycleConfigName

The name of the lifecycle configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

OnCreate

The shell script that runs only once, when you create a notebook instance.

Type: Array of NotebookInstanceLifecycleHook objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

OnStart

The shell script that runs every time you start a notebook instance, including when you create the notebook instance.

Type: Array of NotebookInstanceLifecycleHook objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeOptimizationJob

Service: Amazon SageMaker Service

Provides the properties of the specified optimization job.

Request Syntax

```
{
   "OptimizationJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

OptimizationJobName

The name that you assigned to the optimization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "DeploymentInstanceType": "string",
   "FailureReason": "string",
   "LastModifiedTime": number,
   "MaxInstanceCount": number,
   "ModelSource": {
      "S3": {
         "ModelAccessConfig": {
            "AcceptEula": boolean
         },
```

```
            "S3Uri": "string"
         },
         "SageMakerModel": {
            "ModelName": "string"
         }
      },
      "OptimizationConfigs": [
         { ... }
      ],
      "OptimizationEndTime": number,
      "OptimizationEnvironment": {
         "string" : "string"
      },
      "OptimizationJobArn": "string",
      "OptimizationJobName": "string",
      "OptimizationJobStatus": "string",
      "OptimizationOutput": {
         "RecommendedInferenceImage": "string"
      },
      "OptimizationStartTime": number,
      "OutputConfig": {
         "KmsKeyId": "string",
         "S3OutputLocation": "string",
         "SageMakerModel": {
            "ModelName": "string"
         }
      },
      "RoleArn": "string",
      "StoppingCondition": {
         "MaxPendingTimeInSeconds": number,
         "MaxRuntimeInSeconds": number,
         "MaxWaitTimeInSeconds": number
      },
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The time when you created the optimization job.

Type: Timestamp

DeploymentInstanceType

The type of instance that hosts the optimized model that you create with the optimization job.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge`

FailureReason

If the optimization job status is `FAILED`, the reason for the failure.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

LastModifiedTime

The time when the optimization job was last updated.

Type: Timestamp

MaxInstanceCount

The maximum number of instances to use for the optimization job.

Type: Integer

Valid Range: Minimum value of 1.

ModelSource

The location of the source model to optimize with an optimization job.

Type: OptimizationJobModelSource object

OptimizationConfigs

Settings for each of the optimization techniques that the job applies.

Type: Array of OptimizationConfig objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

OptimizationEndTime

The time when the optimization job finished processing.

Type: Timestamp

OptimizationEnvironment

The environment variables to set in the model container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 25 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `(?!\s*$).+`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

OptimizationJobArn

The Amazon Resource Name (ARN) of the optimization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:optimization-job/.*`

OptimizationJobName

The name that you assigned to the optimization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

OptimizationJobStatus

The current status of the optimization job.

Type: String

Valid Values: `INPROGRESS | COMPLETED | FAILED | STARTING | STOPPING | STOPPED`

OptimizationOutput

Output values produced by an optimization job.

Type: OptimizationOutput object

OptimizationStartTime

The time when the optimization job started.

Type: Timestamp

OutputConfig

Details for where to store the optimized model that you create with the optimization job.

Type: OptimizationJobOutputConfig object

RoleArn

The ARN of the IAM role that you assigned to the optimization job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

StoppingCondition

Specifies a limit to how long a job can run. When the job reaches the time limit, SageMaker ends the job. Use this API to cap costs.

To stop a training job, SageMaker sends the algorithm the `SIGTERM` signal, which delays job termination for 120 seconds. Algorithms can use this 120-second window to save the model artifacts, so the results of training are not lost.

The training algorithms provided by SageMaker automatically save the intermediate results of a model training job when possible. This attempt to save artifacts is only a best effort case as model might not be in a state from which it can be saved. For example, if training has just started, the model might not be ready to save. When saved, this intermediate data is a valid model artifact. You can use it to create a model with `CreateModel`.

> **Note**
>
> The Neural Topic Model (NTM) currently does not support saving intermediate model artifacts. When training NTMs, make sure that the maximum runtime is sufficient for the training job to complete.

Type: StoppingCondition object

VpcConfig

A VPC in Amazon VPC that your optimized model has access to.

Type: OptimizationVpcConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribePartnerApp

Service: Amazon SageMaker Service

Gets information about a SageMaker Partner AI App.

Request Syntax

```
{
   "Arn": "string",
   "IncludeAvailableUpgrade": boolean
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Arn

The ARN of the SageMaker Partner AI App to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:partner-app\/app-[A-Z0-9]{12}`

Required: Yes

IncludeAvailableUpgrade

When set to TRUE, the response includes available upgrade information for the SageMaker Partner AI App. Default is FALSE.

Type: Boolean

Required: No

Response Syntax

```
{
```

```
   "ApplicationConfig": {
      "AdminUsers": [ "string" ],
      "Arguments": {
         "string" : "string"
      },
      "AssignedGroupPatterns": [ "string" ],
      "RoleGroupAssignments": [
         {
            "GroupPatterns": [ "string" ],
            "RoleName": "string"
         }
      ]
   },
   "Arn": "string",
   "AuthType": "string",
   "AvailableUpgrade": {
      "ReleaseNotes": [ "string" ],
      "Version": "string"
   },
   "BaseUrl": "string",
   "CreationTime": number,
   "CurrentVersionEolDate": number,
   "EnableAutoMinorVersionUpgrade": boolean,
   "EnableIamSessionBasedIdentity": boolean,
   "Error": {
      "Code": "string",
      "Reason": "string"
   },
   "ExecutionRoleArn": "string",
   "KmsKeyId": "string",
   "LastModifiedTime": number,
   "MaintenanceConfig": {
      "MaintenanceWindowStart": "string"
   },
   "Name": "string",
   "Status": "string",
   "Tier": "string",
   "Type": "string",
   "Version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ApplicationConfig

Configuration settings for the SageMaker Partner AI App.

Type: PartnerAppConfig object

Arn

The ARN of the SageMaker Partner AI App that was described.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:partner-app\/app-[A-Z0-9]{12}`

AuthType

The authorization type that users use to access the SageMaker Partner AI App.

Type: String

Valid Values: `IAM`

AvailableUpgrade

A map of available minor version upgrades for the SageMaker Partner AI App. The key is the semantic version number, and the value is a list of release notes for that version. A null value indicates no upgrades are available.

Type: AvailableUpgrade object

BaseUrl

The URL of the SageMaker Partner AI App that the Application SDK uses to support in-app calls for the user.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

CreationTime

The time that the SageMaker Partner AI App was created.

Type: Timestamp

CurrentVersionEolDate

The end-of-life date for the current version of the SageMaker Partner AI App.

Type: Timestamp

EnableAutoMinorVersionUpgrade

Indicates whether the SageMaker Partner AI App is configured for automatic minor version upgrades during scheduled maintenance windows.

Type: Boolean

EnableIamSessionBasedIdentity

When set to TRUE, the SageMaker Partner AI App sets the AWS IAM session name or the authenticated IAM user as the identity of the SageMaker Partner AI App user.

Type: Boolean

Error

This is an error field object that contains the error code and the reason for an operation failure.

Type: ErrorInfo object

ExecutionRoleArn

The ARN of the IAM role associated with the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

KmsKeyId

The AWS KMS customer managed key used to encrypt the data at rest associated with SageMaker Partner AI Apps.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

LastModifiedTime

The time that the SageMaker Partner AI App was last modified.

Type: Timestamp

MaintenanceConfig

Maintenance configuration settings for the SageMaker Partner AI App.

Type: PartnerAppMaintenanceConfig object

Name

The name of the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]+`

Status

The status of the SageMaker Partner AI App.

- Creating: SageMaker AI is creating the partner AI app. The partner AI app is not available during creation.
- Updating: SageMaker AI is updating the partner AI app. The partner AI app is not available when updating.
- Deleting: SageMaker AI is deleting the partner AI app. The partner AI app is not available during deletion.
- Available: The partner AI app is provisioned and accessible.
- Failed: The partner AI app is in a failed state and isn't available. SageMaker AI is investigating the issue. For further guidance, contact AWS Support.
- UpdateFailed: The partner AI app couldn't be updated but is available.
- Deleted: The partner AI app is permanently deleted and not available.

Type: String

Valid Values: `Creating | Updating | Deleting | Available | Failed | UpdateFailed | Deleted`

Tier

The instance type and size of the cluster attached to the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `(?!\s*$).+`

Type

The type of SageMaker Partner AI App. Must be one of the following: `lakera-guard`, `comet`, `deepchecks-llm-evaluation`, or `fiddler`.

Type: String

Valid Values: `lakera-guard | comet | deepchecks-llm-evaluation | fiddler`

Version

The version of the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `(?!\s*$).+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribePipeline

Service: Amazon SageMaker Service

Describes the details of a pipeline.

Request Syntax

```
{
   "PipelineName": "string",
   "PipelineVersionId": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

PipelineName

The name or Amazon Resource Name (ARN) of the pipeline to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,255})`

Required: Yes

PipelineVersionId

The ID of the pipeline version to describe.

Type: Long

Valid Range: Minimum value of 1.

Required: No

Response Syntax

```
{
```

```
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "CreationTime": number,
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "LastRunTime": number,
   "ParallelismConfiguration": {
      "MaxParallelExecutionSteps": number
   },
   "PipelineArn": "string",
   "PipelineDefinition": "string",
   "PipelineDescription": "string",
   "PipelineDisplayName": "string",
   "PipelineName": "string",
   "PipelineStatus": "string",
   "PipelineVersionDescription": "string",
   "PipelineVersionDisplayName": "string",
   "RoleArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

The time when the pipeline was created.

Type: Timestamp

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

The time when the pipeline was last modified.

Type: Timestamp

LastRunTime

The time when the pipeline was last run.

Type: Timestamp

ParallelismConfiguration

Lists the parallelism configuration applied to the pipeline.

Type: ParallelismConfiguration object

PipelineArn

The Amazon Resource Name (ARN) of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

PipelineDefinition

The JSON pipeline definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1048576.

Pattern: `.*(?:[ \r\n\t].*)*`

PipelineDescription

The description of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

PipelineDisplayName

The display name of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

PipelineName

The name of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

PipelineStatus

The status of the pipeline execution.

Type: String

Valid Values: `Active | Deleting`

PipelineVersionDescription

The description of the pipeline version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

PipelineVersionDisplayName

The display name of the pipeline version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 82.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,81}`

RoleArn

The Amazon Resource Name (ARN) that the pipeline uses to execute.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribePipelineDefinitionForExecution

Service: Amazon SageMaker Service

Describes the details of an execution's pipeline definition.

Request Syntax

```
{
   "PipelineExecutionArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "PipelineDefinition": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The time when the pipeline was created.

Type: Timestamp

PipelineDefinition

The JSON pipeline definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1048576.

Pattern: `.*(?:[ \r\n\t].*)*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribePipelineExecution

Service: Amazon SageMaker Service

Describes the details of a pipeline execution.

Request Syntax

```
{
   "PipelineExecutionArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: Yes

Response Syntax

```
{
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
```

```
   },
   "CreationTime": number,
   "FailureReason": "string",
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "MLflowConfig": {
      "MlflowExperimentName": "string",
      "MlflowResourceArn": "string"
   },
   "ParallelismConfiguration": {
      "MaxParallelExecutionSteps": number
   },
   "PipelineArn": "string",
   "PipelineExecutionArn": "string",
   "PipelineExecutionDescription": "string",
   "PipelineExecutionDisplayName": "string",
   "PipelineExecutionStatus": "string",
   "PipelineExperimentConfig": {
      "ExperimentName": "string",
      "TrialName": "string"
   },
   "PipelineVersionId": number,
   "SelectiveExecutionConfig": {
      "SelectedSteps": [
         {
            "StepName": "string"
         }
      ],
      "SourcePipelineExecutionArn": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

The time when the pipeline execution was created.

Type: Timestamp

FailureReason

If the execution failed, a message describing why.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1300.

Pattern: `.*`

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

The time when the pipeline execution was modified last.

Type: Timestamp

MLflowConfig

The MLflow configuration of the pipeline execution.

Type: MLflowConfiguration object

ParallelismConfiguration

The parallelism configuration applied to the pipeline.

Type: ParallelismConfiguration object

PipelineArn

The Amazon Resource Name (ARN) of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

PipelineExecutionDescription

The description of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

PipelineExecutionDisplayName

The display name of the pipeline execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 82.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,81}`

PipelineExecutionStatus

The status of the pipeline execution.

Type: String

Valid Values: `Executing | Stopping | Stopped | Failed | Succeeded`

PipelineExperimentConfig

Specifies the names of the experiment and trial created by a pipeline.

Type: PipelineExperimentConfig object

PipelineVersionId

The ID of the pipeline version.

Type: Long

Valid Range: Minimum value of 1.

SelectiveExecutionConfig

The selective execution configuration applied to the pipeline run.

Type: SelectiveExecutionConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeProcessingJob

Service: Amazon SageMaker Service

Returns a description of a processing job.

Request Syntax

```
{
   "ProcessingJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ProcessingJobName

The name of the processing job. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AppSpecification": {
      "ContainerArguments": [ "string" ],
      "ContainerEntrypoint": [ "string" ],
      "ImageUri": "string"
   },
   "AutoMLJobArn": "string",
   "CreationTime": number,
   "Environment": {
      "string" : "string"
```

```
    },
    "ExitMessage": "string",
    "ExperimentConfig": {
       "ExperimentName": "string",
       "RunName": "string",
       "TrialComponentDisplayName": "string",
       "TrialName": "string"
    },
    "FailureReason": "string",
    "LastModifiedTime": number,
    "MonitoringScheduleArn": "string",
    "NetworkConfig": {
       "EnableInterContainerTrafficEncryption": boolean,
       "EnableNetworkIsolation": boolean,
       "VpcConfig": {
          "SecurityGroupIds": [ "string" ],
          "Subnets": [ "string" ]
       }
    },
    "ProcessingEndTime": number,
    "ProcessingInputs": [
       {
          "AppManaged": boolean,
          "DatasetDefinition": {
             "AthenaDatasetDefinition": {
                "Catalog": "string",
                "Database": "string",
                "KmsKeyId": "string",
                "OutputCompression": "string",
                "OutputFormat": "string",
                "OutputS3Uri": "string",
                "QueryString": "string",
                "WorkGroup": "string"
             },
             "DataDistributionType": "string",
             "InputMode": "string",
             "LocalPath": "string",
             "RedshiftDatasetDefinition": {
                "ClusterId": "string",
                "ClusterRoleArn": "string",
                "Database": "string",
                "DbUser": "string",
                "KmsKeyId": "string",
                "OutputCompression": "string",
```

```
            "OutputFormat": "string",
            "OutputS3Uri": "string",
            "QueryString": "string"
         }
      },
      "InputName": "string",
      "S3Input": {
         "LocalPath": "string",
         "S3CompressionType": "string",
         "S3DataDistributionType": "string",
         "S3DataType": "string",
         "S3InputMode": "string",
         "S3Uri": "string"
      }
   }
],
"ProcessingJobArn": "string",
"ProcessingJobName": "string",
"ProcessingJobStatus": "string",
"ProcessingOutputConfig": {
   "KmsKeyId": "string",
   "Outputs": [
      {
         "AppManaged": boolean,
         "FeatureStoreOutput": {
            "FeatureGroupName": "string"
         },
         "OutputName": "string",
         "S3Output": {
            "LocalPath": "string",
            "S3UploadMode": "string",
            "S3Uri": "string"
         }
      }
   ]
},
"ProcessingResources": {
   "ClusterConfig": {
      "InstanceCount": number,
      "InstanceType": "string",
      "VolumeKmsKeyId": "string",
      "VolumeSizeInGB": number
   }
},
```

```
   "ProcessingStartTime": number,
   "RoleArn": "string",
   "StoppingCondition": {
      "MaxRuntimeInSeconds": number
   },
   "TrainingJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AppSpecification

Configures the processing job to run a specified container image.

Type: AppSpecification object

AutoMLJobArn

The ARN of an AutoML job associated with this processing job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

CreationTime

The time at which the processing job was created.

Type: Timestamp

Environment

The environment variables set in the Docker container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[\S\s]*`

ExitMessage

An optional string, up to one KB in size, that contains metadata from the processing container when the processing job exits.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\S\s]*`

ExperimentConfig

The configuration information used to create an experiment.

Type: ExperimentConfig object

FailureReason

A string, up to one KB in size, that contains the reason a processing job failed, if it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

LastModifiedTime

The time at which the processing job was last modified.

Type: Timestamp

MonitoringScheduleArn

The ARN of a monitoring schedule for an endpoint associated with this processing job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

NetworkConfig

Networking options for a processing job.

Type: NetworkConfig object

ProcessingEndTime

The time at which the processing job completed.

Type: Timestamp

ProcessingInputs

The inputs for a processing job.

Type: Array of ProcessingInput objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

ProcessingJobArn

The Amazon Resource Name (ARN) of the processing job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:processing-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ProcessingJobName

The name of the processing job. The name must be unique within an AWS Region in the AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

ProcessingJobStatus

Provides the status of a processing job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

ProcessingOutputConfig

Output configuration for the processing job.

Type: ProcessingOutputConfig object

ProcessingResources

Identifies the resources, ML compute instances, and ML storage volumes to deploy for a processing job. In distributed training, you specify more than one instance.

Type: ProcessingResources object

ProcessingStartTime

The time at which the processing job started.

Type: Timestamp

RoleArn

The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker can assume to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

StoppingCondition

The time limit for how long the processing job is allowed to run.

Type: ProcessingStoppingCondition object

TrainingJobArn

The ARN of a training job associated with this processing job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeProject

Service: Amazon SageMaker Service

Describes the details of a project.

Request Syntax

```
{
   "ProjectName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ProjectName

The name of the project to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

Response Syntax

```
{
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "CreationTime": number,
```

```
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "ProjectArn": "string",
   "ProjectDescription": "string",
   "ProjectId": "string",
   "ProjectName": "string",
   "ProjectStatus": "string",
   "ServiceCatalogProvisionedProductDetails": {
      "ProvisionedProductId": "string",
      "ProvisionedProductStatusMessage": "string"
   },
   "ServiceCatalogProvisioningDetails": {
      "PathId": "string",
      "ProductId": "string",
      "ProvisioningArtifactId": "string",
      "ProvisioningParameters": [
         {
            "Key": "string",
            "Value": "string"
         }
      ]
   },
   "TemplateProviderDetails": [
      {
         "CfnTemplateProviderDetail": {
            "Parameters": [
               {
                  "Key": "string",
                  "Value": "string"
               }
            ],
            "RoleARN": "string",
            "StackDetail": {
               "Id": "string",
               "Name": "string",
```

```
                "StatusMessage": "string"
            },
            "TemplateName": "string",
            "TemplateURL": "string"
         }
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

CreationTime

The time when the project was created.

Type: Timestamp

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

LastModifiedTime

The timestamp when project was last modified.

Type: Timestamp

ProjectArn

The Amazon Resource Name (ARN) of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:project/[\S]{1,2048}`

ProjectDescription

The description of the project.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

ProjectId

The ID of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

ProjectName

The name of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

ProjectStatus

The status of the project.

Type: String

Valid Values: `Pending | CreateInProgress | CreateCompleted | CreateFailed | DeleteInProgress | DeleteFailed | DeleteCompleted | UpdateInProgress | UpdateCompleted | UpdateFailed`

ServiceCatalogProvisionedProductDetails

Information about a provisioned service catalog product.

Type: ServiceCatalogProvisionedProductDetails object

ServiceCatalogProvisioningDetails

Information used to provision a service catalog product. For information, see What is AWS Service Catalog.

Type: ServiceCatalogProvisioningDetails object

TemplateProviderDetails

An array of template providers associated with the project.

Type: Array of TemplateProviderDetail objects

Array Members: Fixed number of 1 item.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeReservedCapacity

Service: Amazon SageMaker Service

Retrieves details about a reserved capacity.

Request Syntax

```
{
   "ReservedCapacityArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ReservedCapacityArn

ARN of the reserved capacity to describe.

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:reserved-capacity/.*`

Required: Yes

Response Syntax

```
{
   "AvailabilityZone": "string",
   "AvailableInstanceCount": number,
   "DurationHours": number,
   "DurationMinutes": number,
   "EndTime": number,
   "InstanceType": "string",
   "InUseInstanceCount": number,
   "ReservedCapacityArn": "string",
```

```
   "ReservedCapacityType": "string",
   "StartTime": number,
   "Status": "string",
   "TotalInstanceCount": number,
   "UltraServerSummary": {
      "AvailableSpareInstanceCount": number,
      "InstanceType": "string",
      "UltraServerCount": number,
      "UltraServerType": "string",
      "UnhealthyInstanceCount": number
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AvailabilityZone

The Availability Zone where the reserved capacity is provisioned.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-z]+\-[0-9a-z\-]+`

AvailableInstanceCount

The number of instances currently available for use in this reserved capacity.

Type: Integer

Valid Range: Minimum value of 0.

DurationHours

The total duration of the reserved capacity in hours.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 87600.

DurationMinutes

The number of minutes for the duration of the reserved capacity. For example, if a reserved capacity starts at 08:55 and ends at 11:30, the minutes field would be 35.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 59.

EndTime

The timestamp when the reserved capacity expires.

Type: Timestamp

InstanceType

The Amazon EC2 instance type used in the reserved capacity.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.trn1.32xlarge | ml.trn2.48xlarge | ml.p6-b200.48xlarge | ml.p4de.24xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge | ml.p6-b300.48xlarge`

InUseInstanceCount

The number of instances currently in use from this reserved capacity.

Type: Integer

Valid Range: Minimum value of 0.

ReservedCapacityArn

ARN of the reserved capacity.

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:reserved-capacity/.*`

ReservedCapacityType

The type of reserved capacity.

Type: String

Valid Values: `UltraServer | Instance`

StartTime

The timestamp when the reserved capacity becomes active.

Type: Timestamp

Status

The current status of the reserved capacity.

Type: String

Valid Values: `Pending | Active | Scheduled | Expired | Failed`

TotalInstanceCount

The total number of instances allocated to this reserved capacity.

Type: Integer

Valid Range: Minimum value of 0.

UltraServerSummary

A summary of the UltraServer associated with this reserved capacity.

Type: UltraServerSummary object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeSpace

Service: Amazon SageMaker Service

Describes the space.

Request Syntax

```
{
   "DomainId": "string",
   "SpaceName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainId

The ID of the associated domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

SpaceName

The name of the space.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "DomainId": "string",
   "FailureReason": "string",
   "HomeEfsFileSystemUid": "string",
   "LastModifiedTime": number,
   "OwnershipSettings": {
      "OwnerUserProfileName": "string"
   },
   "SpaceArn": "string",
   "SpaceDisplayName": "string",
   "SpaceName": "string",
   "SpaceSettings": {
      "AppType": "string",
      "CodeEditorAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number
            }
         },
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         }
      },
      "CustomFileSystems": [
         { ... }
      ],
      "JupyterLabAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number
            }
         },
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
```

```
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         }
      },
      "JupyterServerAppSettings": {
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "KernelGatewayAppSettings": {
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
```

```
      },
      "RemoteAccess": "string",
      "SpaceManagedResources": "string",
      "SpaceStorageSettings": {
         "EbsStorageSettings": {
            "EbsVolumeSizeInGb": number
         }
      }
   },
   "SpaceSharingSettings": {
      "SharingType": "string"
   },
   "Status": "string",
   "Url": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The creation time.

Type: Timestamp

DomainId

The ID of the associated domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

FailureReason

The failure reason.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

HomeEfsFileSystemUid

The ID of the space's profile in the Amazon EFS volume.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `\d+`

LastModifiedTime

The last modified time.

Type: Timestamp

OwnershipSettings

The collection of ownership settings for a space.

Type: OwnershipSettings object

SpaceArn

The space's Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:space/.*`

SpaceDisplayName

The name of the space that appears in the Amazon SageMaker Studio UI.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `(?!\s*$).+`

SpaceName

The name of the space.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

SpaceSettings

A collection of space settings.

Type: SpaceSettings object

SpaceSharingSettings

The collection of space sharing settings for a space.

Type: SpaceSharingSettings object

Status

The status.

Type: String

Valid Values: `Deleting | Failed | InService | Pending | Updating | Update_Failed | Delete_Failed`

Url

Returns the URL of the space. If the space is created with AWS IAM Identity Center (Successor to AWS Single Sign-On) authentication, users can navigate to the URL after appending the respective redirect parameter for the application type to be federated through AWS IAM Identity Center.

The following application types are supported:

- Studio Classic: `&redirect=JupyterServer`
- JupyterLab: `&redirect=JupyterLab`
- Code Editor, based on Code-OSS, Visual Studio Code - Open Source: `&redirect=CodeEditor`

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeStudioLifecycleConfig

Service: Amazon SageMaker Service

Describes the Amazon SageMaker AI Studio Lifecycle Configuration.

Request Syntax

```
{
   "StudioLifecycleConfigName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

StudioLifecycleConfigName

The name of the Amazon SageMaker AI Studio Lifecycle Configuration to describe.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "LastModifiedTime": number,
   "StudioLifecycleConfigAppType": "string",
   "StudioLifecycleConfigArn": "string",
   "StudioLifecycleConfigContent": "string",
   "StudioLifecycleConfigName": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The creation time of the Amazon SageMaker AI Studio Lifecycle Configuration.

Type: Timestamp

LastModifiedTime

This value is equivalent to CreationTime because Amazon SageMaker AI Studio Lifecycle Configurations are immutable.

Type: Timestamp

StudioLifecycleConfigAppType

The App type that the Lifecycle Configuration is attached to.

Type: String

Valid Values: `JupyterServer | KernelGateway | CodeEditor | JupyterLab`

StudioLifecycleConfigArn

The ARN of the Lifecycle Configuration to describe.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:studio-lifecycle-config/.*|None)`

StudioLifecycleConfigContent

The content of your Amazon SageMaker AI Studio Lifecycle Configuration script.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 16384.

Pattern: `[\S\s]+`

StudioLifecycleConfigName

The name of the Amazon SageMaker AI Studio Lifecycle Configuration that is described.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeSubscribedWorkteam

Service: Amazon SageMaker Service

Gets information about a work team provided by a vendor. It returns details about the subscription with a vendor in the AWS Marketplace.

Request Syntax

```
{
   "WorkteamArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

WorkteamArn

The Amazon Resource Name (ARN) of the subscribed work team to describe.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:workteam/.*`

Required: Yes

Response Syntax

```
{
   "SubscribedWorkteam": {
      "ListingId": "string",
      "MarketplaceDescription": "string",
      "MarketplaceTitle": "string",
      "SellerName": "string",
      "WorkteamArn": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

SubscribedWorkteam

A `Workteam` instance that contains information about the work team.

Type: SubscribedWorkteam object

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeTrainingJob

Service: Amazon SageMaker Service

Returns information about a training job.

Some of the attributes below only appear if the training job successfully starts. If the training job fails, `TrainingJobStatus` is `Failed` and, depending on the `FailureReason`, attributes like `TrainingStartTime`, `TrainingTimeInSeconds`, `TrainingEndTime`, and `BillableTimeInSeconds` may not be present in the response.

Request Syntax

```
{
   "TrainingJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrainingJobName

The name of the training job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AlgorithmSpecification": {
      "AlgorithmName": "string",
      "ContainerArguments": [ "string" ],
      "ContainerEntrypoint": [ "string" ],
      "EnableSageMakerMetricsTimeSeries": boolean,
```

```
      "MetricDefinitions": [
         {
            "Name": "string",
            "Regex": "string"
         }
      ],
      "TrainingImage": "string",
      "TrainingImageConfig": {
         "TrainingRepositoryAccessMode": "string",
         "TrainingRepositoryAuthConfig": {
            "TrainingRepositoryCredentialsProviderArn": "string"
         }
      },
      "TrainingInputMode": "string"
   },
   "AutoMLJobArn": "string",
   "BillableTimeInSeconds": number,
   "BillableTokenCount": number,
   "CheckpointConfig": {
      "LocalPath": "string",
      "S3Uri": "string"
   },
   "CreationTime": number,
   "DebugHookConfig": {
      "CollectionConfigurations": [
         {
            "CollectionName": "string",
            "CollectionParameters": {
               "string" : "string"
            }
         }
      ],
      "HookParameters": {
         "string" : "string"
      },
      "LocalPath": "string",
      "S3OutputPath": "string"
   },
   "DebugRuleConfigurations": [
      {
         "InstanceType": "string",
         "LocalPath": "string",
         "RuleConfigurationName": "string",
         "RuleEvaluatorImage": "string",
```

```
         "RuleParameters": {
            "string" : "string"
         },
         "S3OutputPath": "string",
         "VolumeSizeInGB": number
      }
   ],
   "DebugRuleEvaluationStatuses": [
      {
         "LastModifiedTime": number,
         "RuleConfigurationName": "string",
         "RuleEvaluationJobArn": "string",
         "RuleEvaluationStatus": "string",
         "StatusDetails": "string"
      }
   ],
   "EnableInterContainerTrafficEncryption": boolean,
   "EnableManagedSpotTraining": boolean,
   "EnableNetworkIsolation": boolean,
   "Environment": {
      "string" : "string"
   },
   "ExperimentConfig": {
      "ExperimentName": "string",
      "RunName": "string",
      "TrialComponentDisplayName": "string",
      "TrialName": "string"
   },
   "FailureReason": "string",
   "FinalMetricDataList": [
      {
         "MetricName": "string",
         "Timestamp": number,
         "Value": number
      }
   ],
   "HyperParameters": {
      "string" : "string"
   },
   "InfraCheckConfig": {
      "EnableInfraCheck": boolean
   },
   "InputDataConfig": [
      {
```

```
            "ChannelName": "string",
            "CompressionType": "string",
            "ContentType": "string",
            "DataSource": {
               "DatasetSource": {
                  "DatasetArn": "string"
               },
               "FileSystemDataSource": {
                  "DirectoryPath": "string",
                  "FileSystemAccessMode": "string",
                  "FileSystemId": "string",
                  "FileSystemType": "string"
               },
               "S3DataSource": {
                  "AttributeNames": [ "string" ],
                  "HubAccessConfig": {
                     "HubContentArn": "string"
                  },
                  "InstanceGroupNames": [ "string" ],
                  "ModelAccessConfig": {
                     "AcceptEula": boolean
                  },
                  "S3DataDistributionType": "string",
                  "S3DataType": "string",
                  "S3Uri": "string"
               }
            },
            "InputMode": "string",
            "RecordWrapperType": "string",
            "ShuffleConfig": {
               "Seed": number
            }
         }
      ],
      "LabelingJobArn": "string",
      "LastModifiedTime": number,
      "MlflowConfig": {
         "MlflowExperimentName": "string",
         "MlflowResourceArn": "string",
         "MlflowRunName": "string"
      },
      "MlflowDetails": {
         "MlflowExperimentId": "string",
         "MlflowRunId": "string"
```

```
    },
    "ModelArtifacts": {
       "S3ModelArtifacts": "string"
    },
    "ModelPackageConfig": {
       "ModelPackageGroupArn": "string",
       "SourceModelPackageArn": "string"
    },
    "OutputDataConfig": {
       "CompressionType": "string",
       "KmsKeyId": "string",
       "S3OutputPath": "string"
    },
    "OutputModelPackageArn": "string",
    "ProfilerConfig": {
       "DisableProfiler": boolean,
       "ProfilingIntervalInMilliseconds": number,
       "ProfilingParameters": {
          "string" : "string"
       },
       "S3OutputPath": "string"
    },
    "ProfilerRuleConfigurations": [
       {
          "InstanceType": "string",
          "LocalPath": "string",
          "RuleConfigurationName": "string",
          "RuleEvaluatorImage": "string",
          "RuleParameters": {
             "string" : "string"
          },
          "S3OutputPath": "string",
          "VolumeSizeInGB": number
       }
    ],
    "ProfilerRuleEvaluationStatuses": [
       {
          "LastModifiedTime": number,
          "RuleConfigurationName": "string",
          "RuleEvaluationJobArn": "string",
          "RuleEvaluationStatus": "string",
          "StatusDetails": "string"
       }
    ],
```

```
   "ProfilingStatus": "string",
   "ProgressInfo": {
      "CurrentEpoch": number,
      "CurrentStep": number,
      "MaxEpoch": number,
      "TotalStepCountPerEpoch": number
   },
   "RemoteDebugConfig": {
      "EnableRemoteDebug": boolean
   },
   "ResourceConfig": {
      "InstanceCount": number,
      "InstanceGroups": [
         {
            "InstanceCount": number,
            "InstanceGroupName": "string",
            "InstanceType": "string"
         }
      ],
      "InstancePlacementConfig": {
         "EnableMultipleJobs": boolean,
         "PlacementSpecifications": [
            {
               "InstanceCount": number,
               "UltraServerId": "string"
            }
         ]
      },
      "InstanceType": "string",
      "KeepAlivePeriodInSeconds": number,
      "TrainingPlanArn": "string",
      "VolumeKmsKeyId": "string",
      "VolumeSizeInGB": number
   },
   "RetryStrategy": {
      "MaximumRetryAttempts": number
   },
   "RoleArn": "string",
   "SecondaryStatus": "string",
   "SecondaryStatusTransitions": [
      {
         "EndTime": number,
         "StartTime": number,
         "Status": "string",
```

```
            "StatusMessage": "string"
         }
      ],
      "ServerlessJobConfig": {
         "AcceptEula": boolean,
         "BaseModelArn": "string",
         "CustomizationTechnique": "string",
         "EvaluationType": "string",
         "EvaluatorArn": "string",
         "JobType": "string",
         "Peft": "string"
      },
      "StoppingCondition": {
         "MaxPendingTimeInSeconds": number,
         "MaxRuntimeInSeconds": number,
         "MaxWaitTimeInSeconds": number
      },
      "TensorBoardOutputConfig": {
         "LocalPath": "string",
         "S3OutputPath": "string"
      },
      "TrainingEndTime": number,
      "TrainingJobArn": "string",
      "TrainingJobName": "string",
      "TrainingJobStatus": "string",
      "TrainingStartTime": number,
      "TrainingTimeInSeconds": number,
      "TuningJobArn": "string",
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      },
      "WarmPoolStatus": {
         "ResourceRetainedBillableTimeInSeconds": number,
         "ReusedByJob": "string",
         "Status": "string"
      }
   }
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AlgorithmSpecification

Information about the algorithm used for training, and algorithm metadata.

Type: AlgorithmSpecification object

AutoMLJobArn

The Amazon Resource Name (ARN) of an AutoML job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

BillableTimeInSeconds

The billable time in seconds. Billable time refers to the absolute wall-clock time.

Multiply `BillableTimeInSeconds` by the number of instances (`InstanceCount`) in your training cluster to get the total compute time SageMaker bills you if you run distributed training. The formula is as follows: `BillableTimeInSeconds * InstanceCount` .

You can calculate the savings from using managed spot training using the formula `(1 - BillableTimeInSeconds / TrainingTimeInSeconds) * 100`. For example, if `BillableTimeInSeconds` is 100 and `TrainingTimeInSeconds` is 500, the savings is 80%.

Type: Integer

Valid Range: Minimum value of 1.

BillableTokenCount

The billable token count for eligible serverless training jobs.

Type: Long

Valid Range: Minimum value of 0.

CheckpointConfig

Contains information about the output location for managed spot training checkpoint data.

Type: CheckpointConfig object

CreationTime

A timestamp that indicates when the training job was created.

Type: Timestamp

DebugHookConfig

Configuration information for the Amazon SageMaker Debugger hook parameters, metric and tensor collections, and storage paths. To learn more about how to configure the `DebugHookConfig` parameter, see Use the SageMaker and Debugger Configuration API Operations to Create, Update, and Debug Your Training Job.

Type: DebugHookConfig object

DebugRuleConfigurations

Configuration information for Amazon SageMaker Debugger rules for debugging output tensors.

Type: Array of DebugRuleConfiguration objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

DebugRuleEvaluationStatuses

Evaluation status of Amazon SageMaker Debugger rules for debugging on a training job.

Type: Array of DebugRuleEvaluationStatus objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

EnableInterContainerTrafficEncryption

To encrypt all communications between ML compute instances in distributed training, choose `True`. Encryption provides greater security for distributed training, but training might take longer. How long it takes depends on the amount of communication between compute instances, especially if you use a deep learning algorithms in distributed training.

Type: Boolean

EnableManagedSpotTraining

A Boolean indicating whether managed spot training is enabled (`True`) or not (`False`).

Type: Boolean

EnableNetworkIsolation

If you want to allow inbound or outbound network calls, except for calls between peers within a training cluster for distributed training, choose `True`. If you enable network isolation for training jobs that are configured to use a VPC, SageMaker downloads and uploads customer data and model artifacts through the specified VPC, but the training container does not have network access.

Type: Boolean

Environment

The environment variables to set in the Docker container.

> **Important**
>
> Do not include any security-sensitive information including account access IDs, secrets, or tokens in any environment fields. As part of the shared responsibility model, you are responsible for any potential exposure, unauthorized access, or compromise of your sensitive data if caused by security-sensitive information included in the request environment variable or plain text fields.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 512.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 512.

Value Pattern: `[\S\s]*`

ExperimentConfig

Associates a SageMaker job as a trial component with an experiment and trial. Specified when you call the following APIs:

- CreateProcessingJob
- CreateTrainingJob

- CreateTransformJob

Type: ExperimentConfig object

FailureReason

If the training job failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

FinalMetricDataList

A collection of `MetricData` objects that specify the names, values, and dates and times that the training algorithm emitted to Amazon CloudWatch.

Type: Array of MetricData objects

Array Members: Minimum number of 0 items. Maximum number of 40 items.

HyperParameters

Algorithm-specific parameters.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

InfraCheckConfig

Contains information about the infrastructure health check configuration for the training job.

Type: InfraCheckConfig object

InputDataConfig

An array of `Channel` objects that describes each data input channel.

Type: Array of Channel objects

Array Members: Minimum number of 1 item. Maximum number of 20 items.

LabelingJobArn

The Amazon Resource Name (ARN) of the SageMaker Ground Truth labeling job that created the transform or training job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:labeling-job/.*`

LastModifiedTime

A timestamp that indicates when the status of the training job was last modified.

Type: Timestamp

MlflowConfig

The MLflow configuration using SageMaker managed MLflow.

Type: MlflowConfig object

MlflowDetails

The MLflow details of this job.

Type: MlflowDetails object

ModelArtifacts

Information about the Amazon S3 location that is configured for storing model artifacts.

Type: ModelArtifacts object

ModelPackageConfig

The configuration for the model package.

Type: ModelPackageConfig object

OutputDataConfig

The S3 path where model artifacts that you configured when creating the job are stored. SageMaker creates subfolders for model artifacts.

Type: OutputDataConfig object

OutputModelPackageArn

The Amazon Resource Name (ARN) of the output model package containing model weights or checkpoints.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

ProfilerConfig

Configuration information for Amazon SageMaker Debugger system monitoring, framework profiling, and storage paths.

Type: ProfilerConfig object

ProfilerRuleConfigurations

Configuration information for Amazon SageMaker Debugger rules for profiling system and framework metrics.

Type: Array of ProfilerRuleConfiguration objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

ProfilerRuleEvaluationStatuses

Evaluation status of Amazon SageMaker Debugger rules for profiling on a training job.

Type: Array of ProfilerRuleEvaluationStatus objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

ProfilingStatus

Profiling status of a training job.

Type: String

Valid Values: `Enabled | Disabled`

ProgressInfo

The Serverless training job progress information.

Type: TrainingProgressInfo object

RemoteDebugConfig

Configuration for remote debugging. To learn more about the remote debugging functionality of SageMaker, see Access a training container through AWS Systems Manager (SSM) for remote debugging.

Type: RemoteDebugConfig object

ResourceConfig

Resources, including ML compute instances and ML storage volumes, that are configured for model training.

Type: ResourceConfig object

RetryStrategy

The number of times to retry the job when the job fails due to an `InternalServerError`.

Type: RetryStrategy object

RoleArn

The AWS Identity and Access Management (IAM) role configured for the training job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

SecondaryStatus

Provides detailed information about the state of the training job. For detailed information on the secondary status of the training job, see `StatusMessage` under SecondaryStatusTransition.

SageMaker provides primary statuses and secondary statuses that apply to each of them:

InProgress

- `Starting` - Starting the training job.

- `Pending` - The training job is waiting for compute capacity or compute resource provision.
- `Downloading` - An optional stage for algorithms that support `File` training input mode. It indicates that data is being downloaded to the ML storage volumes.
- `Training` - Training is in progress.
- `Interrupted` - The job stopped because the managed spot training instances were interrupted.
- `Uploading` - Training is complete and the model artifacts are being uploaded to the S3 location.

Completed

- `Completed` - The training job has completed.

Failed

- `Failed` - The training job has failed. The reason for the failure is returned in the `FailureReason` field of `DescribeTrainingJobResponse`.

Stopped

- `MaxRuntimeExceeded` - The job stopped because it exceeded the maximum allowed runtime.
- `MaxWaitTimeExceeded` - The job stopped because it exceeded the maximum allowed wait time.
- `Stopped` - The training job has stopped.

Stopping

- `Stopping` - Stopping the training job.

> **Important**
>
> Valid values for `SecondaryStatus` are subject to change.

We no longer support the following secondary statuses:

- `LaunchingMLInstances`
- `PreparingTraining`
- `DownloadingTrainingImage`

Type: String

Valid Values: `Starting | LaunchingMLInstances | PreparingTrainingStack | Downloading | DownloadingTrainingImage | Training | Uploading | Stopping | Stopped | MaxRuntimeExceeded | Completed | Failed | Interrupted | MaxWaitTimeExceeded | Updating | Restarting | Pending`

SecondaryStatusTransitions

A history of all of the secondary statuses that the training job has transitioned through.

Type: Array of SecondaryStatusTransition objects

ServerlessJobConfig

The configuration for serverless training jobs.

Type: ServerlessJobConfig object

StoppingCondition

Specifies a limit to how long a model training job can run. It also specifies how long a managed Spot training job has to complete. When the job reaches the time limit, SageMaker ends the training job. Use this API to cap model training costs.

To stop a job, SageMaker sends the algorithm the `SIGTERM` signal, which delays job termination for 120 seconds. Algorithms can use this 120-second window to save the model artifacts, so the results of training are not lost.

Type: StoppingCondition object

TensorBoardOutputConfig

Configuration of storage locations for the Amazon SageMaker Debugger TensorBoard output data.

Type: TensorBoardOutputConfig object

TrainingEndTime

Indicates the time when the training job ends on training instances. You are billed for the time interval between the value of `TrainingStartTime` and this time. For successful jobs and stopped jobs, this is the time after model artifacts are uploaded. For failed jobs, this is the time when SageMaker detects a job failure.

Type: Timestamp

TrainingJobArn

The Amazon Resource Name (ARN) of the training job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

TrainingJobName

Name of the model training job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

TrainingJobStatus

The status of the training job.

SageMaker provides the following training job statuses:

- `InProgress` - The training is in progress.
- `Completed` - The training job has completed.
- `Failed` - The training job has failed. To see the reason for the failure, see the `FailureReason` field in the response to a `DescribeTrainingJobResponse` call.
- `Stopping` - The training job is stopping.
- `Stopped` - The training job has stopped.

For more detailed information, see `SecondaryStatus`.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting`

TrainingStartTime

Indicates the time when the training job starts on training instances. You are billed for the time interval between this time and the value of `TrainingEndTime`. The start time in CloudWatch

Logs might be later than this time. The difference is due to the time it takes to download the training data and to the size of the training container.

Type: Timestamp

TrainingTimeInSeconds

The training time in seconds.

Type: Integer

Valid Range: Minimum value of 1.

TuningJobArn

The Amazon Resource Name (ARN) of the associated hyperparameter tuning job if the training job was launched by a hyperparameter tuning job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:hyper-parameter-tuning-job/.*`

VpcConfig

A VpcConfig object that specifies the VPC that this training job has access to. For more information, see Protect Training Jobs by Using an Amazon Virtual Private Cloud.

Type: VpcConfig object

WarmPoolStatus

The status of the warm pool associated with the training job.

Type: WarmPoolStatus object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeTrainingPlan

Service: Amazon SageMaker Service

Retrieves detailed information about a specific training plan.

Request Syntax

```
{
   "TrainingPlanName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrainingPlanName

The name of the training plan to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: Yes

Response Syntax

```
{
   "AvailableInstanceCount": number,
   "AvailableSpareInstanceCount": number,
   "CurrencyCode": "string",
   "DurationHours": number,
   "DurationMinutes": number,
   "EndTime": number,
   "InUseInstanceCount": number,
   "ReservedCapacitySummaries": [
      {
```

```
            "AvailabilityZone": "string",
            "AvailabilityZoneId": "string",
            "DurationHours": number,
            "DurationMinutes": number,
            "EndTime": number,
            "InstanceType": "string",
            "ReservedCapacityArn": "string",
            "ReservedCapacityType": "string",
            "StartTime": number,
            "Status": "string",
            "TotalInstanceCount": number,
            "UltraServerCount": number,
            "UltraServerType": "string"
        }
    ],
    "StartTime": number,
    "Status": "string",
    "StatusMessage": "string",
    "TargetResources": [ "string" ],
    "TotalInstanceCount": number,
    "TotalUltraServerCount": number,
    "TrainingPlanArn": "string",
    "TrainingPlanName": "string",
    "UnhealthyInstanceCount": number,
    "UpfrontFee": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AvailableInstanceCount

The number of instances currently available for use in this training plan.

Type: Integer

Valid Range: Minimum value of 0.

AvailableSpareInstanceCount

The number of available spare instances in the training plan.

Type: Integer

Valid Range: Minimum value of 0.

CurrencyCode

The currency code for the upfront fee (e.g., USD).

Type: String

DurationHours

The number of whole hours in the total duration for this training plan.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 87600.

DurationMinutes

The additional minutes beyond whole hours in the total duration for this training plan.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 59.

EndTime

The end time of the training plan.

Type: Timestamp

InUseInstanceCount

The number of instances currently in use from this training plan.

Type: Integer

Valid Range: Minimum value of 0.

ReservedCapacitySummaries

The list of Reserved Capacity providing the underlying compute resources of the plan.

Type: Array of ReservedCapacitySummary objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

StartTime

The start time of the training plan.

Type: Timestamp

Status

The current status of the training plan (e.g., Pending, Active, Expired). To see the complete list of status values available for a training plan, refer to the `Status` attribute within the `TrainingPlanSummary` object.

Type: String

Valid Values: `Pending | Active | Scheduled | Expired | Failed`

StatusMessage

A message providing additional information about the current status of the training plan.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

TargetResources

The target resources (e.g., SageMaker Training Jobs, SageMaker HyperPod, SageMaker Endpoints, Studio apps) that can use this training plan.

Training plans are specific to their target resource.

- A training plan designed for SageMaker training jobs can only be used to schedule and run training jobs.
- A training plan for HyperPod clusters can be used exclusively to provide compute resources to a cluster's instance group.
- A training plan for SageMaker endpoints can be used exclusively to provide compute resources to SageMaker endpoints for model deployment.
- A training plan for Studio apps can be used to launch JupyterLab and Code Editor apps on reserved training plan capacity.

Type: Array of strings

Array Members: Minimum number of 1 item.

Valid Values: `training-job | hyperpod-cluster | endpoint | studio-apps`

TotalInstanceCount

The total number of instances reserved in this training plan.

Type: Integer

Valid Range: Minimum value of 0.

TotalUltraServerCount

The total number of UltraServers reserved to this training plan.

Type: Integer

Valid Range: Minimum value of 1.

TrainingPlanArn

The Amazon Resource Name (ARN); of the training plan.

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

TrainingPlanName

The name of the training plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

UnhealthyInstanceCount

The number of instances in the training plan that are currently in an unhealthy state.

Type: Integer

Valid Range: Minimum value of 0.

UpfrontFee

The upfront fee for the training plan.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeTrainingPlanExtensionHistory

Service: Amazon SageMaker Service

Retrieves the extension history for a specified training plan. The response includes details about each extension, such as the offering ID, start and end dates, status, payment status, and cost information.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string",
   "TrainingPlanArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

The maximum number of extensions to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

A token to continue pagination if more results are available.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

TrainingPlanArn

The Amazon Resource Name (ARN); of the training plan to retrieve extension history for.

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Required: Yes

Response Syntax

```
{
   "NextToken": "string",
   "TrainingPlanExtensions": [
      {
         "AvailabilityZone": "string",
         "AvailabilityZoneId": "string",
         "CurrencyCode": "string",
         "DurationHours": number,
         "EndDate": number,
         "ExtendedAt": number,
         "PaymentStatus": "string",
         "StartDate": number,
         "Status": "string",
         "TrainingPlanExtensionOfferingId": "string",
         "UpfrontFee": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

A token to continue pagination if more results are available.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

TrainingPlanExtensions

A list of extensions for the specified training plan.

Type: Array of TrainingPlanExtension objects

Array Members: Minimum number of 0 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeTransformJob

Service: Amazon SageMaker Service

Returns information about a transform job.

Request Syntax

```
{
   "TransformJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TransformJobName

The name of the transform job that you want to view details of.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AutoMLJobArn": "string",
   "BatchStrategy": "string",
   "CreationTime": number,
   "DataCaptureConfig": {
      "DestinationS3Uri": "string",
      "GenerateInferenceId": boolean,
      "KmsKeyId": "string"
   },
   "DataProcessing": {
      "InputFilter": "string",
      "JoinSource": "string",
```

```
      "OutputFilter": "string"
   },
   "Environment": {
      "string" : "string"
   },
   "ExperimentConfig": {
      "ExperimentName": "string",
      "RunName": "string",
      "TrialComponentDisplayName": "string",
      "TrialName": "string"
   },
   "FailureReason": "string",
   "LabelingJobArn": "string",
   "MaxConcurrentTransforms": number,
   "MaxPayloadInMB": number,
   "ModelClientConfig": {
      "InvocationsMaxRetries": number,
      "InvocationsTimeoutInSeconds": number
   },
   "ModelName": "string",
   "TransformEndTime": number,
   "TransformInput": {
      "CompressionType": "string",
      "ContentType": "string",
      "DataSource": {
         "S3DataSource": {
            "S3DataType": "string",
            "S3Uri": "string"
         }
      },
      "SplitType": "string"
   },
   "TransformJobArn": "string",
   "TransformJobName": "string",
   "TransformJobStatus": "string",
   "TransformOutput": {
      "Accept": "string",
      "AssembleWith": "string",
      "KmsKeyId": "string",
      "S3OutputPath": "string"
   },
   "TransformResources": {
      "InstanceCount": number,
      "InstanceType": "string",
```

```
      "TransformAmiVersion": "string",
      "VolumeKmsKeyId": "string"
   },
   "TransformStartTime": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AutoMLJobArn

The Amazon Resource Name (ARN) of the AutoML transform job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

BatchStrategy

Specifies the number of records to include in a mini-batch for an HTTP inference request. A *record* is a single unit of input data that inference can be made on. For example, a single line in a CSV file is a record.

To enable the batch strategy, you must set `SplitType` to `Line`, `RecordIO`, or `TFRecord`.

Type: String

Valid Values: `MultiRecord | SingleRecord`

CreationTime

A timestamp that shows when the transform Job was created.

Type: Timestamp

DataCaptureConfig

Configuration to control how SageMaker captures inference data.

Type: BatchDataCaptureConfig object

DataProcessing

The data structure used to specify the data to be used for inference in a batch transform job and to associate the data that is relevant to the prediction results in the output. The input filter provided allows you to exclude input data that is not needed for inference in a batch transform job. The output filter provided allows you to include input data relevant to interpreting the predictions in the output from the job. For more information, see Associate Prediction Results with their Corresponding Input Records.

Type: DataProcessing object

Environment

The environment variables to set in the Docker container. We support up to 16 key and values entries in the map.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 16 items.

Key Length Constraints: Minimum length of 0. Maximum length of 1024.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]{0,1023}`

Value Length Constraints: Minimum length of 0. Maximum length of 10240.

Value Pattern: `[\S\s]*`

ExperimentConfig

Associates a SageMaker job as a trial component with an experiment and trial. Specified when you call the following APIs:

- CreateProcessingJob
- CreateTrainingJob
- CreateTransformJob

Type: ExperimentConfig object

FailureReason

If the transform job failed, `FailureReason` describes why it failed. A transform job creates a log file, which includes error messages, and stores it as an Amazon S3 object. For more information, see Log Amazon SageMaker Events with Amazon CloudWatch.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

LabelingJobArn

The Amazon Resource Name (ARN) of the Amazon SageMaker Ground Truth labeling job that created the transform or training job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:labeling-job/.*`

MaxConcurrentTransforms

The maximum number of parallel requests on each instance node that can be launched in a transform job. The default value is 1.

Type: Integer

Valid Range: Minimum value of 0.

MaxPayloadInMB

The maximum payload size, in MB, used in the transform job.

Type: Integer

Valid Range: Minimum value of 0.

ModelClientConfig

The timeout and maximum number of retries for processing a transform job invocation.

Type: ModelClientConfig object

ModelName

The name of the model used in the transform job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

TransformEndTime

Indicates when the transform job has been completed, or has stopped or failed. You are billed for the time interval between this time and the value of `TransformStartTime`.

Type: Timestamp

TransformInput

Describes the dataset to be transformed and the Amazon S3 location where it is stored.

Type: TransformInput object

TransformJobArn

The Amazon Resource Name (ARN) of the transform job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:transform-job/.*`

TransformJobName

The name of the transform job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

TransformJobStatus

The status of the transform job. If the transform job failed, the reason is returned in the `FailureReason` field.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

TransformOutput

Identifies the Amazon S3 location where you want Amazon SageMaker to save the results from the transform job.

Type: TransformOutput object

TransformResources

Describes the resources, including ML instance types and ML instance count, to use for the transform job.

Type: TransformResources object

TransformStartTime

Indicates when the transform job starts on ML instances. You are billed for the time interval between this time and the value of `TransformEndTime`.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeTrial

Service: Amazon SageMaker Service

Provides a list of a trial's properties.

Request Syntax

```
{
   "TrialName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrialName

The name of the trial to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
```

```
   },
   "CreationTime": number,
   "DisplayName": "string",
   "ExperimentName": "string",
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "MetadataProperties": {
      "CommitId": "string",
      "GeneratedBy": "string",
      "ProjectId": "string",
      "Repository": "string"
   },
   "Source": {
      "SourceArn": "string",
      "SourceType": "string"
   },
   "TrialArn": "string",
   "TrialName": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreatedBy

Who created the trial.

Type: UserContext object

CreationTime

When the trial was created.

Type: Timestamp

DisplayName

The name of the trial as displayed. If `DisplayName` isn't specified, `TrialName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

ExperimentName

The name of the experiment the trial is part of.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

LastModifiedBy

Who last modified the trial.

Type: UserContext object

LastModifiedTime

When the trial was last modified.

Type: Timestamp

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Source

The Amazon Resource Name (ARN) of the source and, optionally, the job type.

Type: TrialSource object

TrialArn

The Amazon Resource Name (ARN) of the trial.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial/.*`

TrialName

The name of the trial.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeTrialComponent

Service: Amazon SageMaker Service

Provides a list of a trials component's properties.

Request Syntax

```
{
   "TrialComponentName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrialComponentName

The name of the trial component to describe.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial|experiment-trial-component|artifact|action|context)\/)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,119})`

Required: Yes

Response Syntax

```
{
   "CreatedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
```

```
      "UserProfileName": "string"
   },
   "CreationTime": number,
   "DisplayName": "string",
   "EndTime": number,
   "InputArtifacts": {
      "string" : {
         "MediaType": "string",
         "Value": "string"
      }
   },
   "LastModifiedBy": {
      "DomainId": "string",
      "IamIdentity": {
         "Arn": "string",
         "PrincipalId": "string",
         "SourceIdentity": "string"
      },
      "UserProfileArn": "string",
      "UserProfileName": "string"
   },
   "LastModifiedTime": number,
   "LineageGroupArn": "string",
   "MetadataProperties": {
      "CommitId": "string",
      "GeneratedBy": "string",
      "ProjectId": "string",
      "Repository": "string"
   },
   "Metrics": [
      {
         "Avg": number,
         "Count": number,
         "Last": number,
         "Max": number,
         "MetricName": "string",
         "Min": number,
         "SourceArn": "string",
         "StdDev": number,
         "TimeStamp": number
      }
   ],
   "OutputArtifacts": {
      "string" : {
```

```
            "MediaType": "string",
            "Value": "string"
         }
      },
      "Parameters": {
         "string" : {
            "NumberValue": number,
            "StringValue": "string"
         }
      },
      "Source": {
         "SourceArn": "string",
         "SourceType": "string"
      },
      "Sources": [
         {
            "SourceArn": "string",
            "SourceType": "string"
         }
      ],
      "StartTime": number,
      "Status": {
         "Message": "string",
         "PrimaryStatus": "string"
      },
      "TrialComponentArn": "string",
      "TrialComponentName": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreatedBy

Who created the trial component.

Type: UserContext object

CreationTime

When the component was created.

Type: Timestamp

DisplayName

The name of the component as displayed. If `DisplayName` isn't specified, `TrialComponentName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

EndTime

When the component ended.

Type: Timestamp

InputArtifacts

The input artifacts of the component.

Type: String to TrialComponentArtifact object map

Map Entries: Minimum number of 0 items. Maximum number of 60 items.

Key Length Constraints: Minimum length of 0. Maximum length of 128.

Key Pattern: `.*`

LastModifiedBy

Who last modified the component.

Type: UserContext object

LastModifiedTime

When the component was last modified.

Type: Timestamp

LineageGroupArn

The Amazon Resource Name (ARN) of the lineage group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:lineage-group/.*`

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Metrics

The metrics for the component.

Type: Array of TrialComponentMetricSummary objects

OutputArtifacts

The output artifacts of the component.

Type: String to TrialComponentArtifact object map

Map Entries: Minimum number of 0 items. Maximum number of 60 items.

Key Length Constraints: Minimum length of 0. Maximum length of 128.

Key Pattern: `.*`

Parameters

The hyperparameters of the component.

Type: String to TrialComponentParameterValue object map

Map Entries: Minimum number of 0 items. Maximum number of 300 items.

Key Length Constraints: Minimum length of 0. Maximum length of 320.

Key Pattern: `.*`

Source

The Amazon Resource Name (ARN) of the source and, optionally, the job type.

Type: TrialComponentSource object

Sources

A list of ARNs and, if applicable, job types for multiple sources of an experiment run.

Type: Array of TrialComponentSource objects

StartTime

When the component started.

Type: Timestamp

Status

The status of the component. States include:

- InProgress
- Completed
- Failed

Type: TrialComponentStatus object

TrialComponentArn

The Amazon Resource Name (ARN) of the trial component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial-component/.*`

TrialComponentName

The name of the trial component.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeUserProfile

Service: Amazon SageMaker Service

Describes a user profile. For more information, see `CreateUserProfile`.

Request Syntax

```
{
   "DomainId": "string",
   "UserProfileName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

UserProfileName

The user profile name. This value is not case sensitive.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "CreationTime": number,
   "DomainId": "string",
   "FailureReason": "string",
   "HomeEfsFileSystemUid": "string",
   "LastModifiedTime": number,
   "SingleSignOnUserIdentifier": "string",
   "SingleSignOnUserValue": "string",
   "Status": "string",
   "UserProfileArn": "string",
   "UserProfileName": "string",
   "UserSettings": {
      "AutoMountHomeEFS": "string",
      "CanvasAppSettings": {
         "DirectDeploySettings": {
            "Status": "string"
         },
         "EmrServerlessSettings": {
            "ExecutionRoleArn": "string",
            "Status": "string"
         },
         "GenerativeAiSettings": {
            "AmazonBedrockRoleArn": "string"
         },
         "IdentityProviderOAuthSettings": [
            {
               "DataSourceName": "string",
               "SecretArn": "string",
               "Status": "string"
            }
         ],
         "KendraSettings": {
            "Status": "string"
         },
         "ModelRegisterSettings": {
            "CrossAccountModelRegisterRoleArn": "string",
            "Status": "string"
         },
         "TimeSeriesForecastingSettings": {
            "AmazonForecastRoleArn": "string",
            "Status": "string"
```

```
         },
         "WorkspaceSettings": {
            "S3ArtifactPath": "string",
            "S3KmsKeyId": "string"
         }
      },
      "CodeEditorAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number,
               "LifecycleManagement": "string",
               "MaxIdleTimeoutInMinutes": number,
               "MinIdleTimeoutInMinutes": number
            }
         },
         "BuiltInLifecycleConfigArn": "string",
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "CustomFileSystemConfigs": [
         { ... }
      ],
      "CustomPosixUserConfig": {
         "Gid": number,
         "Uid": number
      },
      "DefaultLandingUri": "string",
      "ExecutionRole": "string",
      "JupyterLabAppSettings": {
         "AppLifecycleManagement": {
```

```
            "IdleSettings": {
               "IdleTimeoutInMinutes": number,
               "LifecycleManagement": "string",
               "MaxIdleTimeoutInMinutes": number,
               "MinIdleTimeoutInMinutes": number
            }
         },
         "BuiltInLifecycleConfigArn": "string",
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "EmrSettings": {
            "AssumableRoleArns": [ "string" ],
            "ExecutionRoleArns": [ "string" ]
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "JupyterServerAppSettings": {
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
```

```
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "KernelGatewayAppSettings": {
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "RSessionAppSettings": {
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         }
      },
      "RStudioServerProAppSettings": {
         "AccessStatus": "string",
         "UserGroup": "string"
```

```
      },
      "SecurityGroups": [ "string" ],
      "SharingSettings": {
         "NotebookOutputOption": "string",
         "S3KmsKeyId": "string",
         "S3OutputPath": "string"
      },
      "SpaceStorageSettings": {
         "DefaultEbsStorageSettings": {
            "DefaultEbsVolumeSizeInGb": number,
            "MaximumEbsVolumeSizeInGb": number
         }
      },
      "StudioWebPortal": "string",
      "StudioWebPortalSettings": {
         "ExecutionRoleSessionNameMode": "string",
         "HiddenAppTypes": [ "string" ],
         "HiddenInstanceTypes": [ "string" ],
         "HiddenMlTools": [ "string" ],
         "HiddenSageMakerImageVersionAliases": [
            {
               "SageMakerImageName": "string",
               "VersionAliases": [ "string" ]
            }
         ]
      },
      "TensorBoardAppSettings": {
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         }
      }
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CreationTime

The creation time.

Type: Timestamp

DomainId

The ID of the domain that contains the profile.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

FailureReason

The failure reason.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

HomeEfsFileSystemUid

The ID of the user's profile in the Amazon Elastic File System volume.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `\d+`

LastModifiedTime

The last modified time.

Type: Timestamp

SingleSignOnUserIdentifier

The IAM Identity Center user identifier.

Type: String

Pattern: `UserName`

SingleSignOnUserValue

The IAM Identity Center user value.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Status

The status.

Type: String

Valid Values: `Deleting | Failed | InService | Pending | Updating | Update_Failed | Delete_Failed`

UserProfileArn

The user profile Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:user-profile/.*`

UserProfileName

The user profile name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

UserSettings

A collection of settings.

Type: UserSettings object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeWorkforce

Service: Amazon SageMaker Service

Lists private workforce information, including workforce name, Amazon Resource Name (ARN), and, if applicable, allowed IP address ranges (CIDRs). Allowable IP address ranges are the IP addresses that workers can use to access tasks.

> **Important**
>
> This operation applies only to private workforces.

Request Syntax

```
{
   "WorkforceName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

WorkforceName

The name of the private workforce whose access you want to restrict. `WorkforceName` is automatically set to `default` when a workforce is created and cannot be modified.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9\-]){0,62}`

Required: Yes

Response Syntax

```
{
```

```
   "Workforce": {
      "CognitoConfig": {
         "ClientId": "string",
         "UserPool": "string"
      },
      "CreateDate": number,
      "FailureReason": "string",
      "IpAddressType": "string",
      "LastUpdatedDate": number,
      "OidcConfig": {
         "AuthenticationRequestExtraParams": {
            "string" : "string"
         },
         "AuthorizationEndpoint": "string",
         "ClientId": "string",
         "Issuer": "string",
         "JwksUri": "string",
         "LogoutEndpoint": "string",
         "Scope": "string",
         "TokenEndpoint": "string",
         "UserInfoEndpoint": "string"
      },
      "SourceIpConfig": {
         "Cidrs": [ "string" ]
      },
      "Status": "string",
      "SubDomain": "string",
      "WorkforceArn": "string",
      "WorkforceName": "string",
      "WorkforceVpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ],
         "VpcEndpointId": "string",
         "VpcId": "string"
      }
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Workforce

A single private workforce, which is automatically created when you create your first private work team. You can create one private work force in each AWS Region. By default, any workforce-related API operation used in a specific region will apply to the workforce created in that region. To learn how to create a private workforce, see Create a Private Workforce.

Type: Workforce object

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeWorkteam

Service: Amazon SageMaker Service

Gets information about a specific work team. You can see information such as the creation date, the last updated date, membership information, and the work team's Amazon Resource Name (ARN).

Request Syntax

```
{
   "WorkteamName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

WorkteamName

The name of the work team to return a description of.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "Workteam": {
      "CreateDate": number,
      "Description": "string",
      "LastUpdatedDate": number,
      "MemberDefinitions": [
         {
            "CognitoMemberDefinition": {
```

```
                "ClientId": "string",
                "UserGroup": "string",
                "UserPool": "string"
             },
             "OidcMemberDefinition": {
                "Groups": [ "string" ]
             }
          }
       ],
       "NotificationConfiguration": {
          "NotificationTopicArn": "string"
       },
       "ProductListingIds": [ "string" ],
       "SubDomain": "string",
       "WorkerAccessConfiguration": {
          "S3Presign": {
             "IamPolicyConstraints": {
                "SourceIp": "string",
                "VpcSourceIp": "string"
             }
          }
       },
       "WorkforceArn": "string",
       "WorkteamArn": "string",
       "WorkteamName": "string"
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Workteam

A `Workteam` instance that contains information about the work team.

Type: Workteam object

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DetachClusterNodeVolume

Service: Amazon SageMaker Service

Detaches your Amazon Elastic Block Store (Amazon EBS) volume from a node in your EKS orchestrated SageMaker HyperPod cluster.

This API works with the Amazon Elastic Block Store (Amazon EBS) Container Storage Interface (CSI) driver to manage the lifecycle of persistent storage in your HyperPod EKS clusters.

Request Syntax

```
{
   "ClusterArn": "string",
   "NodeId": "string",
   "VolumeId": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterArn

The Amazon Resource Name (ARN) of your SageMaker HyperPod cluster containing the target node. Your cluster must use EKS as the orchestration and be in the `InService` state.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Required: Yes

NodeId

The unique identifier of the cluster node from which you want to detach the volume.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: Yes

VolumeId

The unique identifier of your EBS volume that you want to detach. Your volume must be currently attached to the specified node.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `vol-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: Yes

Response Syntax

```
{
   "AttachTime": number,
   "ClusterArn": "string",
   "DeviceName": "string",
   "NodeId": "string",
   "Status": "string",
   "VolumeId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AttachTime

The original timestamp when your volume was initially attached to the node.

Type: Timestamp

ClusterArn

The Amazon Resource Name (ARN) of your SageMaker HyperPod cluster where the volume detachment operation was performed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

DeviceName

The device name assigned to your attached volume on the target instance.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `/dev/(xvd[a-z]|sd[b-z]|sd[f-p][1-6])`

NodeId

The unique identifier of the cluster node from which your volume was detached.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Status

The current status of your volume detachment operation.

Type: String

Valid Values: `attaching | attached | detaching | detached | busy`

VolumeId

The unique identifier of your EBS volume that was detached.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `vol-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DisableSagemakerServicecatalogPortfolio

Service: Amazon SageMaker Service

Disables using Service Catalog in SageMaker. Service Catalog is used to create SageMaker projects.

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DisassociateTrialComponent

Service: Amazon SageMaker Service

Disassociates a trial component from a trial. This doesn't effect other trials the component is associated with. Before you can delete a component, you must disassociate the component from all trials it is associated with. To associate a trial component with a trial, call the AssociateTrialComponent API.

To get a list of the trials a component is associated with, use the Search API. Specify `ExperimentTrialComponent` for the `Resource` parameter. The list appears in the response under `Results.TrialComponent.Parents`.

Request Syntax

```
{
   "TrialComponentName": "string",
   "TrialName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrialComponentName

The name of the component to disassociate from the trial.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

TrialName

The name of the trial to disassociate from.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "TrialArn": "string",
   "TrialComponentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrialArn

The Amazon Resource Name (ARN) of the trial.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial/.*`

TrialComponentArn

The Amazon Resource Name (ARN) of the trial component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial-component/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

EnableSagemakerServicecatalogPortfolio

Service: Amazon SageMaker Service

Enables using Service Catalog in SageMaker. Service Catalog is used to create SageMaker projects.

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ExtendTrainingPlan

Service: Amazon SageMaker Service

Extends an existing training plan by purchasing an extension offering. This allows you to add additional compute capacity time to your training plan without creating a new plan or reconfiguring your workloads.

To find available extension offerings, use the `SearchTrainingPlanOfferings` API with the `TrainingPlanArn` parameter.

To view the history of extensions for a training plan, use the `DescribeTrainingPlanExtensionHistory` API.

Request Syntax

```
{
   "TrainingPlanExtensionOfferingId": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrainingPlanExtensionOfferingId

The unique identifier of the extension offering to purchase. You can retrieve this ID from the `TrainingPlanExtensionOfferings` in the response of the `SearchTrainingPlanOfferings` API.

Type: String

Required: Yes

Response Syntax

```
{
   "TrainingPlanExtensions": [
      {
         "AvailabilityZone": "string",
```

```
            "AvailabilityZoneId": "string",
            "CurrencyCode": "string",
            "DurationHours": number,
            "EndDate": number,
            "ExtendedAt": number,
            "PaymentStatus": "string",
            "StartDate": number,
            "Status": "string",
            "TrainingPlanExtensionOfferingId": "string",
            "UpfrontFee": "string"
        }
    ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrainingPlanExtensions

The list of extensions for the training plan, including the newly created extension.

Type: Array of TrainingPlanExtension objects

Array Members: Minimum number of 0 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetDeviceFleetReport

Service: Amazon SageMaker Service

Describes a fleet.

Request Syntax

```
{
   "DeviceFleetName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeviceFleetName

The name of the fleet.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AgentVersions": [
      {
         "AgentCount": number,
         "Version": "string"
      }
   ],
   "Description": "string",
   "DeviceFleetArn": "string",
   "DeviceFleetName": "string",
```

```
   "DeviceStats": {
      "ConnectedDeviceCount": number,
      "RegisteredDeviceCount": number
   },
   "ModelStats": [
      {
         "ActiveDeviceCount": number,
         "ConnectedDeviceCount": number,
         "ModelName": "string",
         "ModelVersion": "string",
         "OfflineDeviceCount": number,
         "SamplingDeviceCount": number
      }
   ],
   "OutputConfig": {
      "KmsKeyId": "string",
      "PresetDeploymentConfig": "string",
      "PresetDeploymentType": "string",
      "S3OutputLocation": "string"
   },
   "ReportGenerated": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AgentVersions

The versions of Edge Manager agent deployed on the fleet.

Type: Array of AgentVersion objects

Description

Description of the fleet.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 800.

Pattern: `[-a-zA-Z0-9_.,;:! ]*`

DeviceFleetArn

The Amazon Resource Name (ARN) of the device.

Type: String

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:device-fleet/?[a-zA-Z_0-9+=,.@\-_/]+`

DeviceFleetName

The name of the fleet.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

DeviceStats

Status of devices.

Type: DeviceStats object

ModelStats

Status of model on device.

Type: Array of EdgeModelStat objects

OutputConfig

The output configuration for storing sample data collected by the fleet.

Type: EdgeOutputConfig object

ReportGenerated

Timestamp of when the report was generated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetLineageGroupPolicy

Service: Amazon SageMaker Service

The resource policy for the lineage group.

Request Syntax

```
{
   "LineageGroupName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

LineageGroupName

The name or Amazon Resource Name (ARN) of the lineage group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:lineage-group\/)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,119})`

Required: Yes

Response Syntax

```
{
   "LineageGroupArn": "string",
   "ResourcePolicy": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

LineageGroupArn

The Amazon Resource Name (ARN) of the lineage group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:lineage-group/.*`

ResourcePolicy

The resource policy that gives access to the lineage group in another account.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 20480.

Pattern: `.*(?:[ \r\n\t].*)*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetModelPackageGroupPolicy

Service: Amazon SageMaker Service

Gets a resource policy that manages access for a model group. For information about resource policies, see Identity-based policies and resource-based policies in the *AWS Identity and Access Management User Guide.*.

Request Syntax

```
{
   "ModelPackageGroupName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelPackageGroupName

The name of the model group for which to get the resource policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "ResourcePolicy": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ResourcePolicy

The resource policy for the model group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20480.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetSagemakerServicecatalogPortfolioStatus

Service: Amazon SageMaker Service

Gets the status of Service Catalog in SageMaker. Service Catalog is used to create SageMaker projects.

Response Syntax

```
{
   "Status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Status

Whether Service Catalog is enabled or disabled in SageMaker.

Type: String

Valid Values: `Enabled | Disabled`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetScalingConfigurationRecommendation

Service: Amazon SageMaker Service

Starts an Amazon SageMaker Inference Recommender autoscaling recommendation job. Returns recommendations for autoscaling policies that you can apply to your SageMaker endpoint.

Request Syntax

```
{
   "EndpointName": "string",
   "InferenceRecommendationsJobName": "string",
   "RecommendationId": "string",
   "ScalingPolicyObjective": {
      "MaxInvocationsPerMinute": number,
      "MinInvocationsPerMinute": number
   },
   "TargetCpuUtilizationPerCore": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EndpointName

The name of an endpoint benchmarked during a previously completed inference recommendation job. This name should come from one of the recommendations returned by the job specified in the `InferenceRecommendationsJobName` field.

Specify either this field or the `RecommendationId` field.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

InferenceRecommendationsJobName

The name of a previously completed Inference Recommender job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: Yes

RecommendationId

The recommendation ID of a previously completed inference recommendation. This ID should come from one of the recommendations returned by the job specified in the `InferenceRecommendationsJobName` field.

Specify either this field or the `EndpointName` field.

Type: String

Required: No

ScalingPolicyObjective

An object where you specify the anticipated traffic pattern for an endpoint.

Type: ScalingPolicyObjective object

Required: No

TargetCpuUtilizationPerCore

The percentage of how much utilization you want an instance to use before autoscaling. The default value is 50%.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

Response Syntax

```
{
   "DynamicScalingConfiguration": {
```

```
      "MaxCapacity": number,
      "MinCapacity": number,
      "ScaleInCooldown": number,
      "ScaleOutCooldown": number,
      "ScalingPolicies": [
         { ... }
      ]
   },
   "EndpointName": "string",
   "InferenceRecommendationsJobName": "string",
   "Metric": {
      "InvocationsPerInstance": number,
      "ModelLatency": number
   },
   "RecommendationId": "string",
   "ScalingPolicyObjective": {
      "MaxInvocationsPerMinute": number,
      "MinInvocationsPerMinute": number
   },
   "TargetCpuUtilizationPerCore": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

DynamicScalingConfiguration

An object with the recommended values for you to specify when creating an autoscaling policy.

Type: DynamicScalingConfiguration object

EndpointName

The name of an endpoint benchmarked during a previously completed Inference Recommender job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

InferenceRecommendationsJobName

The name of a previously completed Inference Recommender job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Metric

An object with a list of metrics that were benchmarked during the previously completed Inference Recommender job.

Type: ScalingPolicyMetric object

RecommendationId

The recommendation ID of a previously completed inference recommendation.

Type: String

ScalingPolicyObjective

An object representing the anticipated traffic pattern for an endpoint that you specified in the request.

Type: ScalingPolicyObjective object

TargetCpuUtilizationPerCore

The percentage of how much utilization you want an instance to use before autoscaling, which you specified in the request. The default value is 50%.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetSearchSuggestions

Service: Amazon SageMaker Service

An auto-complete API for the search functionality in the SageMaker console. It returns suggestions of possible matches for the property name to use in `Search` queries. Provides suggestions for `HyperParameters`, `Tags`, and `Metrics`.

Request Syntax

```
{
   "Resource": "string",
   "SuggestionQuery": {
      "PropertyNameQuery": {
         "PropertyNameHint": "string"
      }
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Resource

The name of the SageMaker resource to search for.

Type: String

Valid Values: `TrainingJob | Experiment | ExperimentTrial | ExperimentTrialComponent | Endpoint | Model | ModelPackage | ModelPackageGroup | Pipeline | PipelineExecution | FeatureGroup | FeatureMetadata | Image | ImageVersion | Project | HyperParameterTuningJob | ModelCard | PipelineVersion | Job`

Required: Yes

SuggestionQuery

Limits the property names that are included in the response.

Type: SuggestionQuery object

Required: No

Response Syntax

```
{
   "PropertyNameSuggestions": [
      {
         "PropertyName": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

PropertyNameSuggestions

A list of property names for a `Resource` that match a `SuggestionQuery`.

Type: Array of PropertyNameSuggestion objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ImportHubContent

Service: Amazon SageMaker Service

Import hub content.

Request Syntax

```
{
   "DocumentSchemaVersion": "string",
   "HubContentDescription": "string",
   "HubContentDisplayName": "string",
   "HubContentDocument": "string",
   "HubContentMarkdown": "string",
   "HubContentName": "string",
   "HubContentSearchKeywords": [ "string" ],
   "HubContentType": "string",
   "HubContentVersion": "string",
   "HubName": "string",
   "SupportStatus": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DocumentSchemaVersion

The version of the hub content schema to import.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: Yes

HubContentDescription

A description of the hub content to import.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Pattern: `.*`

Required: No

HubContentDisplayName

The display name of the hub content to import.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

HubContentDocument

The hub content document that describes information about the hub content such as type, associated containers, scripts, and more.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 327680.

Pattern: `.*`

Required: Yes

HubContentMarkdown

A string that provides a description of the hub content. This string can include links, tables, and standard markdown formating.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 170391.

Required: No

HubContentName

The name of the hub content to import.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubContentSearchKeywords

The searchable keywords of the hub content.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

HubContentType

The type of hub content to import.

Type: String

Valid Values: `Model | Notebook | ModelReference | DataSet | JsonDoc`

Required: Yes

HubContentVersion

The version of the hub content to import.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: No

HubName

The name of the hub to import content into.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

SupportStatus

The status of the hub content resource.

Type: String

Valid Values: `Supported | Deprecated | Restricted`

Required: No

Tags

Any tags associated with the hub content.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "HubArn": "string",
   "HubContentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HubArn

The ARN of the hub that the content was imported into.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

HubContentArn

The ARN of the hub content that was imported.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListActions

Service: Amazon SageMaker Service

Lists the actions in your account and their properties.

Request Syntax

```
{
   "ActionType": "string",
   "CreatedAfter": number,
   "CreatedBefore": number,
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "SourceUri": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ActionType

A filter that returns only actions of the specified type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

CreatedAfter

A filter that returns only actions created on or after the specified time.

Type: Timestamp

Required: No

CreatedBefore

A filter that returns only actions created on or before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of actions to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous call to `ListActions` didn't return the full set of actions, the call returns a token for getting the next set of actions.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The property used to sort results. The default value is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order. The default value is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

SourceUri

A filter that returns only actions with the specified source URI.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "ActionSummaries": [
      {
         "ActionArn": "string",
         "ActionName": "string",
         "ActionType": "string",
         "CreationTime": number,
         "LastModifiedTime": number,
         "Source": {
            "SourceId": "string",
            "SourceType": "string",
            "SourceUri": "string"
         },
         "Status": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ActionSummaries

A list of actions and their properties.

Type: Array of ActionSummary objects

NextToken

A token for getting the next set of actions, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListAIBenchmarkJobs

Service: Amazon SageMaker Service

Returns a list of AI benchmark jobs in your account. You can filter the results by name, status, and creation time, and sort the results. The response is paginated.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only jobs created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only jobs created before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of benchmark jobs to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the job name. This filter returns only jobs whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the previous call to `ListAIBenchmarkJobs` didn't return the full set of jobs, the call returns a token for getting the next set.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that returns only benchmark jobs with the specified status.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: No

Response Syntax

```
{
   "AIBenchmarkJobs": [
      {
         "AIBenchmarkJobArn": "string",
         "AIBenchmarkJobName": "string",
         "AIBenchmarkJobStatus": "string",
         "AIWorkloadConfigName": "string",
         "CreationTime": number,
         "EndTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIBenchmarkJobs

An array of `AIBenchmarkJobSummary` objects, one for each benchmark job that matches the specified filters.

Type: Array of AIBenchmarkJobSummary objects

NextToken

If the response is truncated, Amazon SageMaker AI returns this token. To retrieve the next set of jobs, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListAIRecommendationJobs

Service: Amazon SageMaker Service

Returns a list of AI recommendation jobs in your account. You can filter the results by name, status, and creation time, and sort the results. The response is paginated.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only jobs created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only jobs created before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of recommendation jobs to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the job name. This filter returns only jobs whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the previous call to `ListAIRecommendationJobs` didn't return the full set of jobs, the call returns a token for getting the next set.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that returns only recommendation jobs with the specified status.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: No

Response Syntax

```
{
   "AIRecommendationJobs": [
      {
         "AIRecommendationJobArn": "string",
         "AIRecommendationJobName": "string",
         "AIRecommendationJobStatus": "string",
         "CreationTime": number,
         "EndTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIRecommendationJobs

An array of `AIRecommendationJobSummary` objects, one for each recommendation job that matches the specified filters.

Type: Array of AIRecommendationJobSummary objects

NextToken

If the response is truncated, Amazon SageMaker AI returns this token. To retrieve the next set of jobs, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListAIWorkloadConfigs

Service: Amazon SageMaker Service

Returns a list of AI workload configurations in your account. You can filter the results by name and creation time, and sort the results. The response is paginated.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only configurations created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only configurations created before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of AI workload configurations to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the configuration name. This filter returns only configurations whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the previous call to `ListAIWorkloadConfigs` didn't return the full set of configurations, the call returns a token for getting the next set of configurations.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order for results. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "AIWorkloadConfigs": [
      {
         "AIWorkloadConfigArn": "string",
         "AIWorkloadConfigName": "string",
         "CreationTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIWorkloadConfigs

An array of `AIWorkloadConfigSummary` objects, one for each AI workload configuration that matches the specified filters.

Type: Array of AIWorkloadConfigSummary objects

NextToken

If the response is truncated, Amazon SageMaker AI returns this token. To retrieve the next set of configurations, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListAlgorithms

Service: Amazon SageMaker Service

Lists the machine learning algorithms that have been created.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only algorithms created after the specified time (timestamp).

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only algorithms created before the specified time (timestamp).

Type: Timestamp

Required: No

MaxResults

The maximum number of algorithms to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the algorithm name. This filter returns only algorithms whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the response to a previous `ListAlgorithms` request was truncated, the response includes a `NextToken`. To retrieve the next set of algorithms, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The parameter by which to sort the results. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order for the results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "AlgorithmSummaryList": [
      {
         "AlgorithmArn": "string",
         "AlgorithmDescription": "string",
         "AlgorithmName": "string",
         "AlgorithmStatus": "string",
         "CreationTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AlgorithmSummaryList

>An array of `AlgorithmSummary` objects, each of which lists an algorithm.

Type: Array of AlgorithmSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of algorithms, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListAliases

Service: Amazon SageMaker Service

Lists the aliases of a specified image or image version.

Request Syntax

```
{
   "Alias": "string",
   "ImageName": "string",
   "MaxResults": number,
   "NextToken": "string",
   "Version": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Alias

The alias of the image version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(?!^[.-])^([a-zA-Z0-9-_.]+)`

Required: No

ImageName

The name of the image.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

MaxResults

The maximum number of aliases to return.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous call to `ListAliases` didn't return the full set of aliases, the call returns a token for retrieving the next set of aliases.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Version

The version of the image. If image version is not specified, the aliases of all versions of the image are listed.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "SageMakerImageVersionAliases": [ "string" ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

A token for getting the next set of aliases, if more aliases exist.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

SageMakerImageVersionAliases

A list of SageMaker AI image version aliases.

Type: Array of strings

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(?!^[.-])^([a-zA-Z0-9-_.]+)`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListAppImageConfigs

Service: Amazon SageMaker Service

Lists the AppImageConfigs in your account and their properties. The list can be filtered by creation time or modified time, and whether the AppImageConfig name contains a specified string.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "ModifiedTimeAfter": number,
   "ModifiedTimeBefore": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only AppImageConfigs created on or after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only AppImageConfigs created on or before the specified time.

Type: Timestamp

Required: No

MaxResults

The total number of items to return in the response. If the total number of items available is more than the value specified, a `NextToken` is provided in the response. To resume pagination, provide the `NextToken` value in the as part of a subsequent call. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

ModifiedTimeAfter

A filter that returns only AppImageConfigs modified on or after the specified time.

Type: Timestamp

Required: No

ModifiedTimeBefore

A filter that returns only AppImageConfigs modified on or before the specified time.

Type: Timestamp

Required: No

NameContains

A filter that returns only AppImageConfigs whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

NextToken

If the previous call to `ListImages` didn't return the full set of AppImageConfigs, the call returns a token for getting the next set of AppImageConfigs.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The property used to sort results. The default value is `CreationTime`.

Type: String

Valid Values: `CreationTime | LastModifiedTime | Name`

Required: No

SortOrder

The sort order. The default value is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "AppImageConfigs": [
      {
         "AppImageConfigArn": "string",
         "AppImageConfigName": "string",
         "CodeEditorAppImageConfig": {
            "ContainerConfig": {
               "ContainerArguments": [ "string" ],
               "ContainerEntrypoint": [ "string" ],
               "ContainerEnvironmentVariables": {
                  "string" : "string"
               }
            },
            "FileSystemConfig": {
               "DefaultGid": number,
               "DefaultUid": number,
               "MountPath": "string"
            }
```

```
         },
         "CreationTime": number,
         "JupyterLabAppImageConfig": {
            "ContainerConfig": {
               "ContainerArguments": [ "string" ],
               "ContainerEntrypoint": [ "string" ],
               "ContainerEnvironmentVariables": {
                  "string" : "string"
               }
            },
            "FileSystemConfig": {
               "DefaultGid": number,
               "DefaultUid": number,
               "MountPath": "string"
            }
         },
         "KernelGatewayImageConfig": {
            "FileSystemConfig": {
               "DefaultGid": number,
               "DefaultUid": number,
               "MountPath": "string"
            },
            "KernelSpecs": [
               {
                  "DisplayName": "string",
                  "Name": "string"
               }
            ]
         },
         "LastModifiedTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AppImageConfigs

A list of AppImageConfigs and their properties.

Type: Array of AppImageConfigDetails objects

NextToken

A token for getting the next set of AppImageConfigs, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListApps

Service: Amazon SageMaker Service

Lists apps.

Request Syntax

```
{
   "DomainIdEquals": "string",
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "SpaceNameEquals": "string",
   "UserProfileNameEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainIdEquals

A parameter to search for the domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: No

MaxResults

This parameter defines the maximum number of results that can be return in a single response. The `MaxResults` parameter is an upper bound, not a target. If there are more results available than the value specified, a `NextToken` is provided in the response. The `NextToken` indicates that the user should get the next set of results by providing this token as a part of a subsequent call. The default value for `MaxResults` is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The parameter by which to sort the results. The default is CreationTime.

Type: String

Valid Values: `CreationTime`

Required: No

SortOrder

The sort order for the results. The default is Ascending.

Type: String

Valid Values: `Ascending | Descending`

Required: No

SpaceNameEquals

A parameter to search by space name. If `UserProfileNameEquals` is set, then this value cannot be set.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

UserProfileNameEquals

A parameter to search by user profile name. If `SpaceNameEquals` is set, then this value cannot be set.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Response Syntax

```
{
   "Apps": [
      {
         "AppName": "string",
         "AppType": "string",
         "CreationTime": number,
         "DomainId": "string",
         "ResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "SpaceName": "string",
         "Status": "string",
         "UserProfileName": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Apps

The list of apps.

Type: Array of AppDetails objects

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

ListArtifacts

Service: Amazon SageMaker Service

Lists the artifacts in your account and their properties.

Request Syntax

```
{
   "ArtifactType": "string",
   "CreatedAfter": number,
   "CreatedBefore": number,
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "SourceUri": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ArtifactType

A filter that returns only artifacts of the specified type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

CreatedAfter

A filter that returns only artifacts created on or after the specified time.

Type: Timestamp

Required: No

CreatedBefore

A filter that returns only artifacts created on or before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of artifacts to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous call to `ListArtifacts` didn't return the full set of artifacts, the call returns a token for getting the next set of artifacts.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The property used to sort results. The default value is `CreationTime`.

Type: String

Valid Values: `CreationTime`

Required: No

SortOrder

The sort order. The default value is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

SourceUri

A filter that returns only artifacts with the specified source URI.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "ArtifactSummaries": [
      {
         "ArtifactArn": "string",
         "ArtifactName": "string",
         "ArtifactType": "string",
         "CreationTime": number,
         "LastModifiedTime": number,
         "Source": {
            "SourceTypes": [
               {
                  "SourceIdType": "string",
                  "Value": "string"
               }
            ],
            "SourceUri": "string"
         }
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ArtifactSummaries

A list of artifacts and their properties.

Type: Array of ArtifactSummary objects

NextToken

A token for getting the next set of artifacts, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

ListAssociations

Service: Amazon SageMaker Service

Lists the associations in your account and their properties.

Request Syntax

```
{
   "AssociationType": "string",
   "CreatedAfter": number,
   "CreatedBefore": number,
   "DestinationArn": "string",
   "DestinationType": "string",
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "SourceArn": "string",
   "SourceType": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AssociationType

A filter that returns only associations of the specified type.

Type: String

Valid Values: `ContributedTo | AssociatedWith | DerivedFrom | Produced | SameAs`

Required: No

CreatedAfter

A filter that returns only associations created on or after the specified time.

Type: Timestamp

Required: No

CreatedBefore

A filter that returns only associations created on or before the specified time.

Type: Timestamp

Required: No

DestinationArn

A filter that returns only associations with the specified destination Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: No

DestinationType

A filter that returns only associations with the specified destination type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

MaxResults

The maximum number of associations to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous call to `ListAssociations` didn't return the full set of associations, the call returns a token for getting the next set of associations.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The property used to sort results. The default value is `CreationTime`.

Type: String

Valid Values: `SourceArn | DestinationArn | SourceType | DestinationType | CreationTime`

Required: No

SortOrder

The sort order. The default value is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

SourceArn

A filter that returns only associations with the specified source ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: No

SourceType

A filter that returns only associations with the specified source type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

Response Syntax

```
{
   "AssociationSummaries": [
      {
         "AssociationType": "string",
         "CreatedBy": {
            "DomainId": "string",
            "IamIdentity": {
               "Arn": "string",
               "PrincipalId": "string",
               "SourceIdentity": "string"
            },
            "UserProfileArn": "string",
            "UserProfileName": "string"
         },
         "CreationTime": number,
         "DestinationArn": "string",
         "DestinationName": "string",
         "DestinationType": "string",
         "SourceArn": "string",
         "SourceName": "string",
         "SourceType": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AssociationSummaries

A list of associations and their properties.

Type: Array of AssociationSummary objects

NextToken

A token for getting the next set of associations, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListAutoMLJobs

Service: Amazon SageMaker Service

Request a list of jobs.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Request a list of jobs, using a filter for time.

Type: Timestamp

Required: No

CreationTimeBefore

Request a list of jobs, using a filter for time.

Type: Timestamp

Required: No

LastModifiedTimeAfter

Request a list of jobs, using a filter for time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

Request a list of jobs, using a filter for time.

Type: Timestamp

Required: No

MaxResults

Request a list of jobs up to a specified limit.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

Request a list of jobs, using a search filter for name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The parameter by which to sort the results. The default is `Name`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for the results. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

Request a list of jobs, using a filter for status.

Type: String

Valid Values: `Completed | InProgress | Failed | Stopped | Stopping`

Required: No

Response Syntax

```
{
   "AutoMLJobSummaries": [
      {
         "AutoMLJobArn": "string",
         "AutoMLJobName": "string",
         "AutoMLJobSecondaryStatus": "string",
         "AutoMLJobStatus": "string",
         "CreationTime": number,
         "EndTime": number,
         "FailureReason": "string",
         "LastModifiedTime": number,
         "PartialFailureReasons": [
```

```
            {
               "PartialFailureMessage": "string"
            }
         ]
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AutoMLJobSummaries

Returns a summary list of jobs.

Type: Array of AutoMLJobSummary objects

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListCandidatesForAutoMLJob

Service: Amazon SageMaker Service

List the candidates created for the job.

Request Syntax

```
{
   "AutoMLJobName": "string",
   "CandidateNameEquals": "string",
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AutoMLJobName

List the candidates created for the job by providing the job's name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

CandidateNameEquals

List the candidates for the job and filter by candidate name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Required: No

MaxResults

List the job's candidates up to a specified limit.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 300.

Required: No

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The parameter by which to sort the results. The default is `Descending`.

Type: String

Valid Values: `CreationTime | Status | FinalObjectiveMetricValue`

Required: No

SortOrder

The sort order for the results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

List the candidates for the job and filter by status.

Type: String

Valid Values: `Completed | InProgress | Failed | Stopped | Stopping`

Required: No

Response Syntax

```
{
   "Candidates": [
      {
         "CandidateName": "string",
         "CandidateProperties": {
            "CandidateArtifactLocations": {
               "BacktestResults": "string",
               "Explainability": "string",
               "ModelInsights": "string"
            },
            "CandidateMetrics": [
               {
                  "MetricName": "string",
                  "Set": "string",
                  "StandardMetricName": "string",
                  "Value": number
               }
            ]
         },
         "CandidateStatus": "string",
         "CandidateSteps": [
            {
               "CandidateStepArn": "string",
               "CandidateStepName": "string",
               "CandidateStepType": "string"
            }
         ],
         "CreationTime": number,
         "EndTime": number,
         "FailureReason": "string",
         "FinalAutoMLJobObjectiveMetric": {
            "MetricName": "string",
            "StandardMetricName": "string",
            "Type": "string",
            "Value": number
```

```
         },
         "InferenceContainerDefinitions": {
            "string" : [
               {
                  "Environment": {
                     "string" : "string"
                  },
                  "Image": "string",
                  "ModelDataUrl": "string"
               }
            ]
         },
         "InferenceContainers": [
            {
               "Environment": {
                  "string" : "string"
               },
               "Image": "string",
               "ModelDataUrl": "string"
            }
         ],
         "LastModifiedTime": number,
         "ObjectiveStatus": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Candidates

Summaries about the `AutoMLCandidates`.

Type: Array of AutoMLCandidate objects

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListClusterEvents

Service: Amazon SageMaker Service

Retrieves a list of event summaries for a specified HyperPod cluster. The operation supports filtering, sorting, and pagination of results. This functionality is only supported when the `NodeProvisioningMode` is set to `Continuous`.

Request Syntax

```
{
   "ClusterName": "string",
   "EventTimeAfter": number,
   "EventTimeBefore": number,
   "InstanceGroupName": "string",
   "MaxResults": number,
   "NextToken": "string",
   "NodeId": "string",
   "ResourceType": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterName

The name or Amazon Resource Name (ARN) of the HyperPod cluster for which to list events.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

EventTimeAfter

The start of the time range for filtering events. Only events that occurred after this time are included in the results.

Type: Timestamp

Required: No

EventTimeBefore

The end of the time range for filtering events. Only events that occurred before this time are included in the results.

Type: Timestamp

Required: No

InstanceGroupName

The name of the instance group to filter events. If specified, only events related to this instance group are returned.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

MaxResults

The maximum number of events to return in the response. Valid range is 1 to 100.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

A token to retrieve the next set of results. This token is obtained from the output of a previous `ListClusterEvents` call.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

NodeId

The EC2 instance ID to filter events. If specified, only events related to this instance are returned.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: No

ResourceType

The type of resource for which to filter events. Valid values are `Cluster`, `InstanceGroup`, or `Instance`.

Type: String

Valid Values: `Cluster | InstanceGroup | Instance`

Required: No

SortBy

The field to use for sorting the event list. Currently, the only supported value is `EventTime`.

Type: String

Valid Values: `EventTime`

Required: No

SortOrder

The order in which to sort the results. Valid values are `Ascending` or `Descending` (the default is `Descending`).

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "Events": [
      {
         "ClusterArn": "string",
         "ClusterName": "string",
         "Description": "string",
         "EventId": "string",
         "EventLevel": "string",
         "EventTime": number,
         "InstanceGroupName": "string",
         "InstanceId": "string",
         "ResourceType": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Events

A list of event summaries matching the specified criteria.

Type: Array of ClusterEventSummary objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

NextToken

A token to retrieve the next set of results. Include this token in subsequent `ListClusterEvents` calls to fetch more events.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListClusterNodes

Service: Amazon SageMaker Service

Retrieves the list of instances (also called *nodes* interchangeably) in a SageMaker HyperPod cluster.

Request Syntax

```
{
   "ClusterName": "string",
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "IncludeNodeLogicalIds": boolean,
   "InstanceGroupNameContains": "string",
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterName

The string name or the Amazon Resource Name (ARN) of the SageMaker HyperPod cluster in which you want to retrieve the list of nodes.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

CreationTimeAfter

A filter that returns nodes in a SageMaker HyperPod cluster created after the specified time. Timestamps are formatted according to the ISO 8601 standard.

Acceptable formats include:

- YYYY-MM-DDThh:mm:ss.sssTZD (UTC), for example, 2014-10-01T20:30:00.000Z
- YYYY-MM-DDThh:mm:ss.sssTZD (with offset), for example, 2014-10-01T12:30:00.000-08:00
- YYYY-MM-DD, for example, 2014-10-01
- Unix time in seconds, for example, 1412195400. This is also referred to as Unix Epoch time and represents the number of seconds since midnight, January 1, 1970 UTC.

For more information about the timestamp format, see Timestamp in the *AWS Command Line Interface User Guide*.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns nodes in a SageMaker HyperPod cluster created before the specified time. The acceptable formats are the same as the timestamp formats for CreationTimeAfter. For more information about the timestamp format, see Timestamp in the *AWS Command Line Interface User Guide*.

Type: Timestamp

Required: No

IncludeNodeLogicalIds

Specifies whether to include nodes that are still being provisioned in the response. When set to true, the response includes all nodes regardless of their provisioning status. When set to False (default), only nodes with assigned InstanceIds are returned.

Type: Boolean

Required: No

InstanceGroupNameContains

A filter that returns the instance groups whose name contain a specified string.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

MaxResults

The maximum number of nodes to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the result of the previous `ListClusterNodes` request was truncated, the response includes a `NextToken`. To retrieve the next set of cluster nodes, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field by which to sort results. The default value is `CREATION_TIME`.

Type: String

Valid Values: `CREATION_TIME | NAME`

Required: No

SortOrder

The sort order for results. The default value is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "ClusterNodeSummaries": [
      {
         "ImageVersionStatus": "string",
         "InstanceGroupName": "string",
         "InstanceId": "string",
         "InstanceStatus": {
            "Message": "string",
            "Status": "string"
         },
         "InstanceType": "string",
         "LastSoftwareUpdateTime": number,
         "LaunchTime": number,
         "NodeLogicalId": "string",
         "PrivateDnsHostname": "string",
         "UltraServerInfo": {
            "Id": "string",
            "Type": "string"
         }
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ClusterNodeSummaries

The summaries of listed instances in a SageMaker HyperPod cluster

Type: Array of ClusterNodeSummary objects

NextToken

The next token specified for listing instances in a SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListClusters

Service: Amazon SageMaker Service

Retrieves the list of SageMaker HyperPod clusters.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "TrainingPlanArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Set a start time for the time range during which you want to list SageMaker HyperPod clusters. Timestamps are formatted according to the ISO 8601 standard.

Acceptable formats include:

- YYYY-MM-DDThh:mm:ss.sssTZD (UTC), for example, 2014-10-01T20:30:00.000Z
- YYYY-MM-DDThh:mm:ss.sssTZD (with offset), for example, 2014-10-01T12:30:00.000-08:00
- YYYY-MM-DD, for example, 2014-10-01
- Unix time in seconds, for example, 1412195400. This is also referred to as Unix Epoch time and represents the number of seconds since midnight, January 1, 1970 UTC.

For more information about the timestamp format, see Timestamp in the *AWS Command Line Interface User Guide*.

Type: Timestamp

Required: No

CreationTimeBefore

Set an end time for the time range during which you want to list SageMaker HyperPod clusters. A filter that returns nodes in a SageMaker HyperPod cluster created before the specified time. The acceptable formats are the same as the timestamp formats for `CreationTimeAfter`. For more information about the timestamp format, see Timestamp in the *AWS Command Line Interface User Guide*.

Type: Timestamp

Required: No

MaxResults

Specifies the maximum number of clusters to evaluate for the operation (not necessarily the number of matching items). After SageMaker processes the number of clusters up to `MaxResults`, it stops the operation and returns the matching clusters up to that point. If all the matching clusters are desired, SageMaker will go through all the clusters until `NextToken` is empty.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

Set the maximum number of instances to print in the list.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

Set the next token to retrieve the list of SageMaker HyperPod clusters.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field by which to sort results. The default value is `CREATION_TIME`.

Type: String

Valid Values: `CREATION_TIME | NAME`

Required: No

SortOrder

The sort order for results. The default value is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

TrainingPlanArn

The Amazon Resource Name (ARN); of the training plan to filter clusters by. For more information about reserving GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Required: No

Response Syntax

```
{
```

```
   "ClusterSummaries": [
      {
         "ClusterArn": "string",
         "ClusterName": "string",
         "ClusterStatus": "string",
         "CreationTime": number,
         "TrainingPlanArns": [ "string" ]
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ClusterSummaries

The summaries of listed SageMaker HyperPod clusters.

Type: Array of ClusterSummary objects

NextToken

If the result of the previous `ListClusters` request was truncated, the response includes a `NextToken`. To retrieve the next set of clusters, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListClusterSchedulerConfigs

Service: Amazon SageMaker Service

List the cluster policy configurations.

Request Syntax

```
{
   "ClusterArn": "string",
   "CreatedAfter": number,
   "CreatedBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "Status": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterArn

Filter for ARN of the cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Required: No

CreatedAfter

Filter for after this creation time. The input for this parameter is a Unix timestamp. To convert a date and time into a Unix timestamp, see EpochConverter.

Type: Timestamp

Required: No

CreatedBefore

Filter for before this creation time. The input for this parameter is a Unix timestamp. To convert a date and time into a Unix timestamp, see EpochConverter.

Type: Timestamp

Required: No

MaxResults

The maximum number of cluster policies to list.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

Filter for name containing this string.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Filter for sorting the list by a given value. For example, sort by name, creation time, or status.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The order of the list. By default, listed in `Descending` order according to by `SortBy`. To change the list order, you can specify `SortOrder` to be `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Status

Filter for status.

Type: String

Valid Values: `Creating | CreateFailed | CreateRollbackFailed | Created | Updating | UpdateFailed | UpdateRollbackFailed | Updated | Deleting | DeleteFailed | DeleteRollbackFailed | Deleted`

Required: No

Response Syntax

```
{
   "ClusterSchedulerConfigSummaries": [
      {
         "ClusterArn": "string",
         "ClusterSchedulerConfigArn": "string",
         "ClusterSchedulerConfigId": "string",
         "ClusterSchedulerConfigVersion": number,
```

```
            "CreationTime": number,
            "LastModifiedTime": number,
            "Name": "string",
            "Status": "string"
        }
    ],
    "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ClusterSchedulerConfigSummaries

Summaries of the cluster policies.

Type: Array of ClusterSchedulerConfigSummary objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListCodeRepositories

Service: Amazon SageMaker Service

Gets a list of the Git repositories in your account.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only Git repositories that were created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only Git repositories that were created before the specified time.

Type: Timestamp

Required: No

LastModifiedTimeAfter

A filter that returns only Git repositories that were last modified after the specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only Git repositories that were last modified before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of Git repositories to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the Git repositories name. This filter returns only repositories whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9-]+`

Required: No

NextToken

If the result of a `ListCodeRepositoriesOutput` request was truncated, the response includes a `NextToken`. To get the next set of Git repositories, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `Name`.

Type: String

Valid Values: `Name | CreationTime | LastModifiedTime`

Required: No

SortOrder

The sort order for results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "CodeRepositorySummaryList": [
      {
         "CodeRepositoryArn": "string",
         "CodeRepositoryName": "string",
         "CreationTime": number,
         "GitConfig": {
            "Branch": "string",
            "RepositoryUrl": "string",
            "SecretArn": "string"
         },
         "LastModifiedTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CodeRepositorySummaryList

Gets a list of summaries of the Git repositories. Each summary specifies the following values for the repository:

- Name
- Amazon Resource Name (ARN)
- Creation time
- Last modified time
- Configuration information, including the URL location of the repository and the ARN of the AWS Secrets Manager secret that contains the credentials used to access the repository.

Type: Array of CodeRepositorySummary objects

NextToken

If the result of a `ListCodeRepositoriesOutput` request was truncated, the response includes a `NextToken`. To get the next set of Git repositories, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListCompilationJobs

Service: Amazon SageMaker Service

Lists model compilation jobs that satisfy various filters.

To create a model compilation job, use CreateCompilationJob. To get information about a particular model compilation job you have created, use DescribeCompilationJob.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns the model compilation jobs that were created after a specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns the model compilation jobs that were created before a specified time.

Type: Timestamp

Required: No

LastModifiedTimeAfter

A filter that returns the model compilation jobs that were modified after a specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns the model compilation jobs that were modified before a specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of model compilation jobs to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A filter that returns the model compilation jobs whose name contains a specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the result of the previous `ListCompilationJobs` request was truncated, the response includes a `NextToken`. To retrieve the next set of model compilation jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field by which to sort results. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that retrieves model compilation jobs with a specific `CompilationJobStatus` status.

Type: String

Valid Values: `INPROGRESS | COMPLETED | FAILED | STARTING | STOPPING | STOPPED`

Required: No

Response Syntax

```
{
   "CompilationJobSummaries": [
```

```
      {
         "CompilationEndTime": number,
         "CompilationJobArn": "string",
         "CompilationJobName": "string",
         "CompilationJobStatus": "string",
         "CompilationStartTime": number,
         "CompilationTargetDevice": "string",
         "CompilationTargetPlatformAccelerator": "string",
         "CompilationTargetPlatformArch": "string",
         "CompilationTargetPlatformOs": "string",
         "CreationTime": number,
         "LastModifiedTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CompilationJobSummaries

An array of CompilationJobSummary objects, each describing a model compilation job.

Type: Array of CompilationJobSummary objects

NextToken

If the response is truncated, Amazon SageMaker AI returns this `NextToken`. To retrieve the next set of model compilation jobs, use this token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListComputeQuotas

Service: Amazon SageMaker Service

List the resource allocation definitions.

Request Syntax

```
{
   "ClusterArn": "string",
   "CreatedAfter": number,
   "CreatedBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "Status": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterArn

Filter for ARN of the cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Required: No

CreatedAfter

Filter for after this creation time. The input for this parameter is a Unix timestamp. To convert a date and time into a Unix timestamp, see EpochConverter.

Type: Timestamp

Required: No

CreatedBefore

Filter for before this creation time. The input for this parameter is a Unix timestamp. To convert a date and time into a Unix timestamp, see EpochConverter.

Type: Timestamp

Required: No

MaxResults

The maximum number of compute allocation definitions to list.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

Filter for name containing this string.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Filter for sorting the list by a given value. For example, sort by name, creation time, or status.

Type: String

Valid Values: `Name | CreationTime | Status | ClusterArn`

Required: No

SortOrder

The order of the list. By default, listed in `Descending` order according to by `SortBy`. To change the list order, you can specify `SortOrder` to be `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Status

Filter for status.

Type: String

Valid Values: `Creating | CreateFailed | CreateRollbackFailed | Created | Updating | UpdateFailed | UpdateRollbackFailed | Updated | Deleting | DeleteFailed | DeleteRollbackFailed | Deleted`

Required: No

Response Syntax

```
{
   "ComputeQuotaSummaries": [
      {
         "ActivationState": "string",
         "ClusterArn": "string",
         "ComputeQuotaArn": "string",
         "ComputeQuotaConfig": {
            "ComputeQuotaResources": [
               {
                  "AcceleratorPartition": {
                     "Count": number,
                     "Type": "string"
                  },
                  "Accelerators": number,
```

```
                "Count": number,
                "InstanceType": "string",
                "MemoryInGiB": number,
                "VCpu": number
             }
          ],
          "PreemptTeamTasks": "string",
          "ResourceSharingConfig": {
             "AbsoluteBorrowLimits": [
                {
                   "AcceleratorPartition": {
                      "Count": number,
                      "Type": "string"
                   },
                   "Accelerators": number,
                   "Count": number,
                   "InstanceType": "string",
                   "MemoryInGiB": number,
                   "VCpu": number
                }
             ],
             "BorrowLimit": number,
             "Strategy": "string"
          }
       },
       "ComputeQuotaId": "string",
       "ComputeQuotaTarget": {
          "FairShareWeight": number,
          "TeamName": "string"
       },
       "ComputeQuotaVersion": number,
       "CreationTime": number,
       "LastModifiedTime": number,
       "Name": "string",
       "Status": "string"
    }
 ],
 "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ComputeQuotaSummaries

Summaries of the compute allocation definitions.

Type: Array of ComputeQuotaSummary objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python

- AWS SDK for Ruby V3

ListContexts

Service: Amazon SageMaker Service

Lists the contexts in your account and their properties.

Request Syntax

```
{
   "ContextType": "string",
   "CreatedAfter": number,
   "CreatedBefore": number,
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "SourceUri": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ContextType

A filter that returns only contexts of the specified type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

CreatedAfter

A filter that returns only contexts created on or after the specified time.

Type: Timestamp

Required: No

CreatedBefore

A filter that returns only contexts created on or before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of contexts to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous call to `ListContexts` didn't return the full set of contexts, the call returns a token for getting the next set of contexts.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The property used to sort results. The default value is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order. The default value is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

SourceUri

A filter that returns only contexts with the specified source URI.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "ContextSummaries": [
      {
         "ContextArn": "string",
         "ContextName": "string",
         "ContextType": "string",
         "CreationTime": number,
         "LastModifiedTime": number,
         "Source": {
            "SourceId": "string",
            "SourceType": "string",
            "SourceUri": "string"
         }
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ContextSummaries

A list of contexts and their properties.

Type: Array of ContextSummary objects

NextToken

A token for getting the next set of contexts, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListDataQualityJobDefinitions

Service: Amazon SageMaker Service

Lists the data quality job definitions in your account.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "EndpointName": "string",
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only data quality monitoring job definitions created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only data quality monitoring job definitions created before the specified time.

Type: Timestamp

Required: No

EndpointName

A filter that lists the data quality job definitions associated with the specified endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MaxResults

The maximum number of data quality monitoring job definitions to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the data quality monitoring job definition name. This filter returns only data quality monitoring job definitions whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the result of the previous `ListDataQualityJobDefinitions` request was truncated, the response includes a `NextToken`. To retrieve the next set of transform jobs, use the token in the next request.>

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

Whether to sort the results in `Ascending` or `Descending` order. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "JobDefinitionSummaries": [
      {
         "CreationTime": number,
         "EndpointName": "string",
         "MonitoringJobDefinitionArn": "string",
         "MonitoringJobDefinitionName": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobDefinitionSummaries

A list of data quality monitoring job definitions.

Type: Array of MonitoringJobDefinitionSummary objects

NextToken

If the result of the previous `ListDataQualityJobDefinitions` request was truncated, the response includes a `NextToken`. To retrieve the next set of data quality monitoring job definitions, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListDeviceFleets

Service: Amazon SageMaker Service

Returns a list of devices in the fleet.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Filter fleets where packaging job was created after specified time.

Type: Timestamp

Required: No

CreationTimeBefore

Filter fleets where the edge packaging job was created before specified time.

Type: Timestamp

Required: No

LastModifiedTimeAfter

Select fleets where the job was updated after X

Type: Timestamp

Required: No

LastModifiedTimeBefore

Select fleets where the job was updated before X

Type: Timestamp

Required: No

MaxResults

The maximum number of results to select.

Type: Integer

Valid Range: Maximum value of 100.

Required: No

NameContains

Filter for fleets containing this name in their fleet device name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

The response from the last list when returning a list large enough to need tokening.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The column to sort by.

Type: String

Valid Values: `NAME | CREATION_TIME | LAST_MODIFIED_TIME`

Required: No

SortOrder

What direction to sort in.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "DeviceFleetSummaries": [
      {
         "CreationTime": number,
         "DeviceFleetArn": "string",
         "DeviceFleetName": "string",
         "LastModifiedTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

DeviceFleetSummaries

Summary of the device fleet.

Type: Array of DeviceFleetSummary objects

NextToken

The response from the last list when returning a list large enough to need tokening.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListDevices

Service: Amazon SageMaker Service

A list of devices.

Request Syntax

```
{
   "DeviceFleetName": "string",
   "LatestHeartbeatAfter": number,
   "MaxResults": number,
   "ModelName": "string",
   "NextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeviceFleetName

Filter for fleets containing this name in their device fleet name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

LatestHeartbeatAfter

Select fleets where the job was updated after X

Type: Timestamp

Required: No

MaxResults

Maximum number of results to select.

Type: Integer

Valid Range: Maximum value of 100.

Required: No

ModelName

A filter that searches devices that contains this name in any of their models.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

NextToken

The response from the last list when returning a list large enough to need tokening.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "DeviceSummaries": [
      {
         "AgentVersion": "string",
         "Description": "string",
         "DeviceArn": "string",
         "DeviceFleetName": "string",
         "DeviceName": "string",
         "IotThingName": "string",
         "LatestHeartbeat": number,
         "Models": [
            {
```

```
                "ModelName": "string",
                "ModelVersion": "string"
            }
         ],
         "RegistrationTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

DeviceSummaries

Summary of devices.

Type: Array of DeviceSummary objects

NextToken

The response from the last list when returning a list large enough to need tokening.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListDomains

Service: Amazon SageMaker Service

Lists the domains.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

This parameter defines the maximum number of results that can be return in a single response. The `MaxResults` parameter is an upper bound, not a target. If there are more results available than the value specified, a `NextToken` is provided in the response. The `NextToken` indicates that the user should get the next set of results by providing this token as a part of a subsequent call. The default value for `MaxResults` is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "Domains": [
      {
         "CreationTime": number,
         "DomainArn": "string",
         "DomainId": "string",
         "DomainName": "string",
         "LastModifiedTime": number,
         "Status": "string",
         "Url": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Domains

The list of domains.

Type: Array of DomainDetails objects

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListEdgeDeploymentPlans

Service: Amazon SageMaker Service

Lists all edge deployment plans.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "DeviceFleetNameContains": "string",
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Selects edge deployment plans created after this time.

Type: Timestamp

Required: No

CreationTimeBefore

Selects edge deployment plans created before this time.

Type: Timestamp

Required: No

DeviceFleetNameContains

Selects edge deployment plans with a device fleet name containing this name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

LastModifiedTimeAfter

Selects edge deployment plans that were last updated after this time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

Selects edge deployment plans that were last updated before this time.

Type: Timestamp

Required: No

MaxResults

The maximum number of results to select (50 by default).

Type: Integer

Valid Range: Maximum value of 100.

Required: No

NameContains

Selects edge deployment plans with names containing this name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

The response from the last list when returning a list large enough to need tokening.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The column by which to sort the edge deployment plans. Can be one of `NAME`, `DEVICEFLEETNAME`, `CREATIONTIME`, `LASTMODIFIEDTIME`.

Type: String

Valid Values: `NAME | DEVICE_FLEET_NAME | CREATION_TIME | LAST_MODIFIED_TIME`

Required: No

SortOrder

The direction of the sorting (ascending or descending).

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "EdgeDeploymentPlanSummaries": [
      {
         "CreationTime": number,
         "DeviceFleetName": "string",
         "EdgeDeploymentFailed": number,
         "EdgeDeploymentPending": number,
         "EdgeDeploymentPlanArn": "string",
         "EdgeDeploymentPlanName": "string",
```

```
            "EdgeDeploymentSuccess": number,
            "LastModifiedTime": number
        }
    ],
    "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

EdgeDeploymentPlanSummaries

List of summaries of edge deployment plans.

Type: Array of EdgeDeploymentPlanSummary objects

NextToken

The token to use when calling the next page of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListEdgePackagingJobs

Service: Amazon SageMaker Service

Returns a list of edge packaging jobs.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "ModelNameContains": "string",
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Select jobs where the job was created after specified time.

Type: Timestamp

Required: No

CreationTimeBefore

Select jobs where the job was created before specified time.

Type: Timestamp

Required: No

LastModifiedTimeAfter

Select jobs where the job was updated after specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

Select jobs where the job was updated before specified time.

Type: Timestamp

Required: No

MaxResults

Maximum number of results to select.

Type: Integer

Valid Range: Maximum value of 100.

Required: No

ModelNameContains

Filter for jobs where the model name contains this string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NameContains

Filter for jobs containing this name in their packaging job name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

The response from the last list when returning a list large enough to need tokening.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Use to specify what column to sort by.

Type: String

Valid Values: `NAME | MODEL_NAME | CREATION_TIME | LAST_MODIFIED_TIME | STATUS`

Required: No

SortOrder

What direction to sort by.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

The job status to filter for.

Type: String

Valid Values: `STARTING | INPROGRESS | COMPLETED | FAILED | STOPPING | STOPPED`

Required: No

Response Syntax

```
{
   "EdgePackagingJobSummaries": [
      {
         "CompilationJobName": "string",
         "CreationTime": number,
         "EdgePackagingJobArn": "string",
         "EdgePackagingJobName": "string",
         "EdgePackagingJobStatus": "string",
         "LastModifiedTime": number,
         "ModelName": "string",
         "ModelVersion": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

EdgePackagingJobSummaries

Summaries of edge packaging jobs.

Type: Array of EdgePackagingJobSummary objects

NextToken

Token to use when calling the next page of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListEndpointConfigs

Service: Amazon SageMaker Service

Lists endpoint configurations.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only endpoint configurations with a creation time greater than or equal to the specified time (timestamp).

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only endpoint configurations created before the specified time (timestamp).

Type: Timestamp

Required: No

MaxResults

The maximum number of training jobs to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the endpoint configuration name. This filter returns only endpoint configurations whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9-]+`

Required: No

NextToken

If the result of the previous `ListEndpointConfig` request was truncated, the response includes a `NextToken`. To retrieve the next set of endpoint configurations, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order for results. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "EndpointConfigs": [
      {
         "CreationTime": number,
         "EndpointConfigArn": "string",
         "EndpointConfigName": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

EndpointConfigs

An array of endpoint configurations.

Type: Array of EndpointConfigSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of endpoint configurations, use it in the subsequent request

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListEndpoints

Service: Amazon SageMaker Service

Lists endpoints.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only endpoints with a creation time greater than or equal to the specified time (timestamp).

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only endpoints that were created before the specified time (timestamp).

Type: Timestamp

Required: No

LastModifiedTimeAfter

A filter that returns only endpoints that were modified after the specified timestamp.

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only endpoints that were modified before the specified timestamp.

Type: Timestamp

Required: No

MaxResults

The maximum number of endpoints to return in the response. This value defaults to 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in endpoint names. This filter returns only endpoints whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9-]+`

Required: No

NextToken

If the result of a `ListEndpoints` request was truncated, the response includes a `NextToken`. To retrieve the next set of endpoints, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Sorts the list of results. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that returns only endpoints with the specified status.

Type: String

Valid Values: `OutOfService | Creating | Updating | SystemUpdating | RollingBack | InService | Deleting | Failed | UpdateRollbackFailed`

Required: No

Response Syntax

```
{
   "Endpoints": [
      {
         "CreationTime": number,
         "EndpointArn": "string",
```

```
            "EndpointName": "string",
            "EndpointStatus": "string",
            "LastModifiedTime": number
        }
    ],
    "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Endpoints

An array or endpoint objects.

Type: Array of EndpointSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of training jobs, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListExperiments

Service: Amazon SageMaker Service

Lists all the experiments in your account. The list can be filtered to show only experiments that were created in a specific time range. The list can be sorted by experiment name or creation time.

Request Syntax

```
{
   "CreatedAfter": number,
   "CreatedBefore": number,
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreatedAfter

A filter that returns only experiments created after the specified time.

Type: Timestamp

Required: No

CreatedBefore

A filter that returns only experiments created before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of experiments to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous call to `ListExperiments` didn't return the full set of experiments, the call returns a token for getting the next set of experiments.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The property used to sort results. The default value is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order. The default value is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "ExperimentSummaries": [
      {
         "CreationTime": number,
         "DisplayName": "string",
         "ExperimentArn": "string",
```

```
            "ExperimentName": "string",
            "ExperimentSource": {
               "SourceArn": "string",
               "SourceType": "string"
            },
            "LastModifiedTime": number
         }
      ],
      "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ExperimentSummaries

A list of the summaries of your experiments.

Type: Array of ExperimentSummary objects

NextToken

A token for getting the next set of experiments, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListFeatureGroups

Service: Amazon SageMaker Service

List `FeatureGroups` based on given filter and order.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "FeatureGroupStatusEquals": "string",
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "OfflineStoreStatusEquals": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Use this parameter to search for `FeatureGroupss` created after a specific date and time.

Type: Timestamp

Required: No

CreationTimeBefore

Use this parameter to search for `FeatureGroupss` created before a specific date and time.

Type: Timestamp

Required: No

FeatureGroupStatusEquals

A `FeatureGroup` status. Filters by `FeatureGroup` status.

Type: String

Valid Values: `Creating | Created | CreateFailed | Deleting | DeleteFailed`

Required: No

MaxResults

The maximum number of results returned by `ListFeatureGroups`.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string that partially matches one or more `FeatureGroups` names. Filters `FeatureGroups` by name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Required: No

NextToken

A token to resume pagination of `ListFeatureGroups` results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

OfflineStoreStatusEquals

An `OfflineStore` status. Filters by `OfflineStore` status.

Type: String

Valid Values: `Active | Blocked | Disabled`

Required: No

SortBy

The value on which the feature group list is sorted.

Type: String

Valid Values: `Name | FeatureGroupStatus | OfflineStoreStatus | CreationTime`

Required: No

SortOrder

The order in which feature groups are listed.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "FeatureGroupSummaries": [
      {
         "CreationTime": number,
         "FeatureGroupArn": "string",
         "FeatureGroupName": "string",
         "FeatureGroupStatus": "string",
         "OfflineStoreStatus": {
            "BlockedReason": "string",
            "Status": "string"
         }
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

FeatureGroupSummaries

A summary of feature groups.

Type: Array of FeatureGroupSummary objects

NextToken

A token to resume pagination of `ListFeatureGroups` results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListFlowDefinitions

Service: Amazon SageMaker Service

Returns information about the flow definitions in your account.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "NextToken": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only flow definitions with a creation time greater than or equal to the specified timestamp.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only flow definitions that were created before the specified timestamp.

Type: Timestamp

Required: No

MaxResults

The total number of items to return. If the total number of available items is more than the value specified in `MaxResults`, then a `NextToken` will be provided in the output that you can use to resume pagination.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

A token to resume pagination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortOrder

An optional value that specifies whether you want the results sorted in `Ascending` or `Descending` order.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "FlowDefinitionSummaries": [
      {
         "CreationTime": number,
         "FailureReason": "string",
         "FlowDefinitionArn": "string",
         "FlowDefinitionName": "string",
         "FlowDefinitionStatus": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

FlowDefinitionSummaries

An array of objects describing the flow definitions.

Type: Array of FlowDefinitionSummary objects

NextToken

A token to resume pagination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python

- AWS SDK for Ruby V3

ListHubContents

Service: Amazon SageMaker Service

List the contents of a hub.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "HubContentType": "string",
   "HubName": "string",
   "MaxResults": number,
   "MaxSchemaVersion": "string",
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Only list hub content that was created after the time specified.

Type: Timestamp

Required: No

CreationTimeBefore

Only list hub content that was created before the time specified.

Type: Timestamp

Required: No

HubContentType

The type of hub content to list.

Type: String

Valid Values: `Model | Notebook | ModelReference | DataSet | JsonDoc`

Required: Yes

HubName

The name of the hub to list the contents of.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

MaxResults

The maximum amount of hub content to list.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

MaxSchemaVersion

The upper bound of the hub content schema verion.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: No

NameContains

Only list hub content if the name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the response to a previous `ListHubContents` request was truncated, the response includes a `NextToken`. To retrieve the next set of hub content, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Sort hub content versions by either name or creation time.

Type: String

Valid Values: `HubContentName | CreationTime | HubContentStatus`

Required: No

SortOrder

Sort hubs by ascending or descending order.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "HubContentSummaries": [
      {
         "CreationTime": number,
```

```
            "DocumentSchemaVersion": "string",
            "HubContentArn": "string",
            "HubContentDescription": "string",
            "HubContentDisplayName": "string",
            "HubContentName": "string",
            "HubContentSearchKeywords": [ "string" ],
            "HubContentStatus": "string",
            "HubContentType": "string",
            "HubContentVersion": "string",
            "OriginalCreationTime": number,
            "SageMakerPublicHubContentArn": "string",
            "SupportStatus": "string"
        }
    ],
    "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HubContentSummaries

The summaries of the listed hub content.

Type: Array of HubContentInfo objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of hub content, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListHubContentVersions

Service: Amazon SageMaker Service

List hub content versions.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "HubContentName": "string",
   "HubContentType": "string",
   "HubName": "string",
   "MaxResults": number,
   "MaxSchemaVersion": "string",
   "MinVersion": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Only list hub content versions that were created after the time specified.

Type: Timestamp

Required: No

CreationTimeBefore

Only list hub content versions that were created before the time specified.

Type: Timestamp

Required: No

HubContentName

The name of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubContentType

The type of hub content to list versions of.

Type: String

Valid Values: `Model | Notebook | ModelReference | DataSet | JsonDoc`

Required: Yes

HubName

The name of the hub to list the content versions of.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

MaxResults

The maximum number of hub content versions to list.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

MaxSchemaVersion

The upper bound of the hub content schema version.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: No

MinVersion

The lower bound of the hub content versions to list.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: No

NextToken

If the response to a previous `ListHubContentVersions` request was truncated, the response includes a `NextToken`. To retrieve the next set of hub content versions, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Sort hub content versions by either name or creation time.

Type: String

Valid Values: `HubContentName | CreationTime | HubContentStatus`

Required: No

SortOrder

Sort hub content versions by ascending or descending order.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "HubContentSummaries": [
      {
         "CreationTime": number,
         "DocumentSchemaVersion": "string",
         "HubContentArn": "string",
         "HubContentDescription": "string",
         "HubContentDisplayName": "string",
         "HubContentName": "string",
         "HubContentSearchKeywords": [ "string" ],
         "HubContentStatus": "string",
         "HubContentType": "string",
         "HubContentVersion": "string",
         "OriginalCreationTime": number,
         "SageMakerPublicHubContentArn": "string",
         "SupportStatus": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HubContentSummaries

The summaries of the listed hub content versions.

Type: Array of HubContentInfo objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of hub content versions, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListHubs

Service: Amazon SageMaker Service

List all existing hubs.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Only list hubs that were created after the time specified.

Type: Timestamp

Required: No

CreationTimeBefore

Only list hubs that were created before the time specified.

Type: Timestamp

Required: No

LastModifiedTimeAfter

Only list hubs that were last modified after the time specified.

Type: Timestamp

Required: No

LastModifiedTimeBefore

Only list hubs that were last modified before the time specified.

Type: Timestamp

Required: No

MaxResults

The maximum number of hubs to list.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

Only list hubs with names that contain the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the response to a previous `ListHubs` request was truncated, the response includes a `NextToken`. To retrieve the next set of hubs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Sort hubs by either name or creation time.

Type: String

Valid Values: `HubName | CreationTime | HubStatus | AccountIdOwner`

Required: No

SortOrder

Sort hubs by ascending or descending order.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "HubSummaries": [
      {
         "CreationTime": number,
         "HubArn": "string",
         "HubDescription": "string",
         "HubDisplayName": "string",
         "HubName": "string",
         "HubSearchKeywords": [ "string" ],
         "HubStatus": "string",
         "LastModifiedTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HubSummaries

The summaries of the listed hubs.

Type: Array of HubInfo objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of hubs, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListHumanTaskUis

Service: Amazon SageMaker Service

Returns information about the human task user interfaces in your account.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "NextToken": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only human task user interfaces with a creation time greater than or equal to the specified timestamp.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only human task user interfaces that were created before the specified timestamp.

Type: Timestamp

Required: No

MaxResults

The total number of items to return. If the total number of available items is more than the value specified in `MaxResults`, then a `NextToken` will be provided in the output that you can use to resume pagination.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

A token to resume pagination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortOrder

An optional value that specifies whether you want the results sorted in `Ascending` or `Descending` order.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "HumanTaskUiSummaries": [
      {
         "CreationTime": number,
         "HumanTaskUiArn": "string",
         "HumanTaskUiName": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HumanTaskUiSummaries

An array of objects describing the human task user interfaces.

Type: Array of HumanTaskUiSummary objects

NextToken

A token to resume pagination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListHyperParameterTuningJobs

Service: Amazon SageMaker Service

Gets a list of HyperParameterTuningJobSummary objects that describe the hyperparameter tuning jobs launched in your account.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only tuning jobs that were created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only tuning jobs that were created before the specified time.

Type: Timestamp

Required: No

LastModifiedTimeAfter

A filter that returns only tuning jobs that were modified after the specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only tuning jobs that were modified before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of tuning jobs to return. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the tuning job name. This filter returns only tuning jobs whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the result of the previous `ListHyperParameterTuningJobs` request was truncated, the response includes a `NextToken`. To retrieve the next set of tuning jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `Name`.

Type: String

Valid Values: `Name | Status | CreationTime`

Required: No

SortOrder

The sort order for results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that returns only tuning jobs with the specified status.

Type: String

Valid Values: `Completed | InProgress | Failed | Stopped | Stopping | Deleting | DeleteFailed`

Required: No

Response Syntax

```
{
   "HyperParameterTuningJobSummaries": [
      {
         "CreationTime": number,
         "HyperParameterTuningEndTime": number,
```

```
         "HyperParameterTuningJobArn": "string",
         "HyperParameterTuningJobName": "string",
         "HyperParameterTuningJobStatus": "string",
         "LastModifiedTime": number,
         "ObjectiveStatusCounters": {
            "Failed": number,
            "Pending": number,
            "Succeeded": number
         },
         "ResourceLimits": {
            "MaxNumberOfTrainingJobs": number,
            "MaxParallelTrainingJobs": number,
            "MaxRuntimeInSeconds": number
         },
         "Strategy": "string",
         "TrainingJobStatusCounters": {
            "Completed": number,
            "InProgress": number,
            "NonRetryableError": number,
            "RetryableError": number,
            "Stopped": number
         }
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HyperParameterTuningJobSummaries

A list of HyperParameterTuningJobSummary objects that describe the tuning jobs that the `ListHyperParameterTuningJobs` request returned.

Type: Array of HyperParameterTuningJobSummary objects

NextToken

If the result of this `ListHyperParameterTuningJobs` request was truncated, the response includes a `NextToken`. To retrieve the next set of tuning jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListImages

Service: Amazon SageMaker Service

Lists the images in your account and their properties. The list can be filtered by creation time or modified time, and whether the image name contains a specified string.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only images created on or after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only images created on or before the specified time.

Type: Timestamp

Required: No

LastModifiedTimeAfter

A filter that returns only images modified on or after the specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only images modified on or before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of images to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A filter that returns only images whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-.]+`

Required: No

NextToken

If the previous call to `ListImages` didn't return the full set of images, the call returns a token for getting the next set of images.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The property used to sort results. The default value is CREATION_TIME.

Type: String

Valid Values: CREATION_TIME | LAST_MODIFIED_TIME | IMAGE_NAME

Required: No

SortOrder

The sort order. The default value is DESCENDING.

Type: String

Valid Values: ASCENDING | DESCENDING

Required: No

Response Syntax

```
{
   "Images": [
      {
         "CreationTime": number,
         "Description": "string",
         "DisplayName": "string",
         "FailureReason": "string",
         "ImageArn": "string",
         "ImageName": "string",
         "ImageStatus": "string",
         "LastModifiedTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Images

A list of images and their properties.

Type: Array of Image objects

NextToken

A token for getting the next set of images, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListImageVersions

Service: Amazon SageMaker Service

Lists the versions of a specified image and their properties. The list can be filtered by creation time or modified time.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "ImageName": "string",
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only versions created on or after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only versions created on or before the specified time.

Type: Timestamp

Required: No

ImageName

The name of the image to list the versions of.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

LastModifiedTimeAfter

A filter that returns only versions modified on or after the specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only versions modified on or before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of versions to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous call to `ListImageVersions` didn't return the full set of versions, the call returns a token for getting the next set of versions.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The property used to sort results. The default value is CREATION_TIME.

Type: String

Valid Values: CREATION_TIME | LAST_MODIFIED_TIME | VERSION

Required: No

SortOrder

The sort order. The default value is DESCENDING.

Type: String

Valid Values: ASCENDING | DESCENDING

Required: No

Response Syntax

```
{
   "ImageVersions": [
      {
         "CreationTime": number,
         "FailureReason": "string",
         "ImageArn": "string",
         "ImageVersionArn": "string",
         "ImageVersionStatus": "string",
         "LastModifiedTime": number,
         "Version": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ImageVersions

A list of versions and their properties.

Type: Array of ImageVersion objects

NextToken

A token for getting the next set of versions, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

ListInferenceComponents

Service: Amazon SageMaker Service

Lists the inference components in your account and their properties.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "EndpointNameEquals": "string",
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string",
   "VariantNameEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Filters the results to only those inference components that were created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

Filters the results to only those inference components that were created before the specified time.

Type: Timestamp

Required: No

EndpointNameEquals

An endpoint name to filter the listed inference components. The response includes only those inference components that are hosted at the specified endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

LastModifiedTimeAfter

Filters the results to only those inference components that were updated after the specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

Filters the results to only those inference components that were updated before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of inference components to return in the response. This value defaults to 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

Filters the results to only those inference components with a name that contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9-]+`

Required: No

NextToken

A token that you use to get the next set of results following a truncated response. If the response to the previous request was truncated, that response provides the value for this token.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field by which to sort the inference components in the response. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

Filters the results to only those inference components with the specified status.

Type: String

Valid Values: `InService | Creating | Updating | Failed | Deleting`

Required: No

VariantNameEquals

A production variant name to filter the listed inference components. The response includes only those inference components that are hosted at the specified variant.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Response Syntax

```
{
   "InferenceComponents": [
      {
         "CreationTime": number,
         "EndpointArn": "string",
         "EndpointName": "string",
         "InferenceComponentArn": "string",
         "InferenceComponentName": "string",
         "InferenceComponentStatus": "string",
         "LastModifiedTime": number,
         "VariantName": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

InferenceComponents

A list of inference components and their properties that matches any of the filters you specified in the request.

Type: Array of InferenceComponentSummary objects

NextToken

The token to use in a subsequent request to get the next set of results following a truncated response.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListInferenceExperiments

Service: Amazon SageMaker Service

Returns the list of all inference experiments.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string",
   "Type": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Selects inference experiments which were created after this timestamp.

Type: Timestamp

Required: No

CreationTimeBefore

Selects inference experiments which were created before this timestamp.

Type: Timestamp

Required: No

LastModifiedTimeAfter

Selects inference experiments which were last modified after this timestamp.

Type: Timestamp

Required: No

LastModifiedTimeBefore

Selects inference experiments which were last modified before this timestamp.

Type: Timestamp

Required: No

MaxResults

The maximum number of results to select.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

Selects inference experiments whose names contain this name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

The response from the last list when returning a list large enough to need tokening.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The column by which to sort the listed inference experiments.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The direction of sorting (ascending or descending).

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

Selects inference experiments which are in this status. For the possible statuses, see DescribeInferenceExperiment.

Type: String

Valid Values: `Creating | Created | Updating | Running | Starting | Stopping | Completed | Cancelled`

Required: No

Type

Selects inference experiments of this type. For the possible types of inference experiments, see CreateInferenceExperiment.

Type: String

Valid Values: `ShadowMode`

Required: No

Response Syntax

```
{
   "InferenceExperiments": [
      {
         "CompletionTime": number,
         "CreationTime": number,
         "Description": "string",
         "LastModifiedTime": number,
         "Name": "string",
         "RoleArn": "string",
         "Schedule": {
            "EndTime": number,
            "StartTime": number
         },
         "Status": "string",
         "StatusReason": "string",
         "Type": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

InferenceExperiments

List of inference experiments.

Type: Array of InferenceExperimentSummary objects

NextToken

The token to use when calling the next page of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListInferenceRecommendationsJobs

Service: Amazon SageMaker Service

Lists recommendation jobs that satisfy various filters.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "ModelNameEquals": "string",
   "ModelPackageVersionArnEquals": "string",
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only jobs created after the specified time (timestamp).

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only jobs created before the specified time (timestamp).

Type: Timestamp

Required: No

LastModifiedTimeAfter

A filter that returns only jobs that were last modified after the specified time (timestamp).

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only jobs that were last modified before the specified time (timestamp).

Type: Timestamp

Required: No

MaxResults

The maximum number of recommendations to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

ModelNameEquals

A filter that returns only jobs that were created for this model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

ModelPackageVersionArnEquals

A filter that returns only jobs that were created for this versioned model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: No

NameContains

A string in the job name. This filter returns only recommendations whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the response to a previous `ListInferenceRecommendationsJobsRequest` request was truncated, the response includes a `NextToken`. To retrieve the next set of recommendations, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The parameter by which to sort the results.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for the results.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that retrieves only inference recommendations jobs with a specific status.

Type: String

Valid Values: `PENDING | IN_PROGRESS | COMPLETED | FAILED | STOPPING | STOPPED | DELETING | DELETED`

Required: No

Response Syntax

```
{
   "InferenceRecommendationsJobs": [
      {
         "CompletionTime": number,
         "CreationTime": number,
         "FailureReason": "string",
         "JobArn": "string",
         "JobDescription": "string",
         "JobName": "string",
         "JobType": "string",
         "LastModifiedTime": number,
         "ModelName": "string",
         "ModelPackageVersionArn": "string",
         "RoleArn": "string",
         "SamplePayloadUrl": "string",
         "Status": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

InferenceRecommendationsJobs

The recommendations created from the Amazon SageMaker Inference Recommender job.

Type: Array of InferenceRecommendationsJob objects

NextToken

A token for getting the next set of recommendations, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListInferenceRecommendationsJobSteps

Service: Amazon SageMaker Service

Returns a list of the subtasks for an Inference Recommender job.

The supported subtasks are benchmarks, which evaluate the performance of your model on different instance types.

Request Syntax

```
{
   "JobName": "string",
   "MaxResults": number,
   "NextToken": "string",
   "Status": "string",
   "StepType": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobName

The name for the Inference Recommender job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: Yes

MaxResults

The maximum number of results to return.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

A token that you can specify to return more results from the list. Specify this field if you have a token that was returned from a previous request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Status

A filter to return benchmarks of a specified status. If this field is left empty, then all benchmarks are returned.

Type: String

Valid Values: `PENDING | IN_PROGRESS | COMPLETED | FAILED | STOPPING | STOPPED | DELETING | DELETED`

Required: No

StepType

A filter to return details about the specified type of subtask.

`BENCHMARK`: Evaluate the performance of your model on different instance types.

Type: String

Valid Values: `BENCHMARK`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "Steps": [
      {
         "InferenceBenchmark": {
            "EndpointConfiguration": {
```

```
            "EndpointName": "string",
            "InitialInstanceCount": number,
            "InstanceType": "string",
            "ServerlessConfig": {
               "MaxConcurrency": number,
               "MemorySizeInMB": number,
               "ProvisionedConcurrency": number
            },
            "VariantName": "string"
         },
         "EndpointMetrics": {
            "MaxInvocations": number,
            "ModelLatency": number
         },
         "FailureReason": "string",
         "InvocationEndTime": number,
         "InvocationStartTime": number,
         "Metrics": {
            "CostPerHour": number,
            "CostPerInference": number,
            "CpuUtilization": number,
            "MaxInvocations": number,
            "MemoryUtilization": number,
            "ModelLatency": number,
            "ModelSetupTime": number
         },
         "ModelConfiguration": {
            "CompilationJobName": "string",
            "EnvironmentParameters": [
               {
                  "Key": "string",
                  "Value": "string",
                  "ValueType": "string"
               }
            ],
            "InferenceSpecificationName": "string"
         }
      },
      "JobName": "string",
      "Status": "string",
      "StepType": "string"
   }
]
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

A token that you can specify in your next request to return more results from the list.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Steps

A list of all subtask details in Inference Recommender.

Type: Array of InferenceRecommendationsJobStep objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListJobs

Service: Amazon SageMaker Service

Lists jobs in a specified category. You can filter results by creation time, last modified time, name, and status. Results are sorted by the field you specify in `SortBy`. Use pagination to retrieve large result sets efficiently.

The following operations are related to `ListJobs`:

- `CreateJob`
- `DescribeJob`

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "JobCategory": "string",
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only jobs created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only jobs created before the specified time.

Type: Timestamp

Required: No

JobCategory

The category of jobs to list.

Type: String

Valid Values: `AgentRFT | AgentRFTEvaluation`

Required: Yes

LastModifiedTimeAfter

A filter that returns only jobs modified after the specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only jobs modified before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of jobs to return in the response. The default value is 50.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the job name to filter results. Only jobs whose name contains the specified string are returned.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the previous response was truncated, this token retrieves the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results. Valid values are `Ascending` and `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that returns only jobs with the specified status.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting | DeleteFailed`

Required: No

Response Syntax

```
{
   "JobSummaries": [
      {
         "CreationTime": number,
         "EndTime": number,
         "JobArn": "string",
         "JobCategory": "string",
         "JobName": "string",
         "JobSecondaryStatus": "string",
         "JobStatus": "string",
         "LastModifiedTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobSummaries

An array of `JobSummary` objects that provide summary information about the jobs.

Type: Array of JobSummary objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

NextToken

If the response is truncated, this token retrieves the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListJobSchemaVersions

Service: Amazon SageMaker Service

Lists available configuration schema versions for a specified job category. Use the schema versions with `DescribeJobSchemaVersion` to retrieve the full schema document.

The following operations are related to `ListJobSchemaVersions`:

- `DescribeJobSchemaVersion`
- `CreateJob`

Request Syntax

```
{
   "JobCategory": "string",
   "MaxResults": number,
   "NextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobCategory

The category of job schemas to list.

Type: String

Valid Values: `AgentRFT | AgentRFTEvaluation`

Required: Yes

MaxResults

The maximum number of schema versions to return in the response. The default value is 5.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous response was truncated, this token retrieves the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "JobConfigSchemas": [
      {
         "JobConfigSchemaVersion": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobConfigSchemas

An array of `JobConfigSchemaVersionSummary` objects listing the available schema versions.

Type: Array of JobConfigSchemaVersionSummary objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

NextToken

If the response is truncated, this token retrieves the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListLabelingJobs

Service: Amazon SageMaker Service

Gets a list of labeling jobs.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only labeling jobs created after the specified time (timestamp).

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only labeling jobs created before the specified time (timestamp).

Type: Timestamp

Required: No

LastModifiedTimeAfter

A filter that returns only labeling jobs modified after the specified time (timestamp).

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only labeling jobs modified before the specified time (timestamp).

Type: Timestamp

Required: No

MaxResults

The maximum number of labeling jobs to return in each page of the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the labeling job name. This filter returns only labeling jobs whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the result of the previous `ListLabelingJobs` request was truncated, the response includes a `NextToken`. To retrieve the next set of labeling jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that retrieves only labeling jobs with a specific status.

Type: String

Valid Values: `Initializing | InProgress | Completed | Failed | Stopping | Stopped`

Required: No

Response Syntax

```
{
   "LabelingJobSummaryList": [
      {
         "AnnotationConsolidationLambdaArn": "string",
         "CreationTime": number,
         "FailureReason": "string",
         "InputConfig": {
            "DataAttributes": {
```

```
               "ContentClassifiers": [ "string" ]
            },
            "DataSource": {
               "S3DataSource": {
                  "ManifestS3Uri": "string"
               },
               "SnsDataSource": {
                  "SnsTopicArn": "string"
               }
            }
         },
         "LabelCounters": {
            "FailedNonRetryableError": number,
            "HumanLabeled": number,
            "MachineLabeled": number,
            "TotalLabeled": number,
            "Unlabeled": number
         },
         "LabelingJobArn": "string",
         "LabelingJobName": "string",
         "LabelingJobOutput": {
            "FinalActiveLearningModelArn": "string",
            "OutputDatasetS3Uri": "string"
         },
         "LabelingJobStatus": "string",
         "LastModifiedTime": number,
         "PreHumanTaskLambdaArn": "string",
         "WorkteamArn": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

LabelingJobSummaryList

An array of `LabelingJobSummary` objects, each describing a labeling job.

Type: Array of LabelingJobSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of labeling jobs, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListLabelingJobsForWorkteam

Service: Amazon SageMaker Service

Gets a list of labeling jobs assigned to a specified work team.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "JobReferenceCodeContains": "string",
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "WorkteamArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only labeling jobs created after the specified time (timestamp).

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only labeling jobs created before the specified time (timestamp).

Type: Timestamp

Required: No

JobReferenceCodeContains

A filter the limits jobs to only the ones whose job reference code contains the specified string.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: No

MaxResults

The maximum number of labeling jobs to return in each page of the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the result of the previous `ListLabelingJobsForWorkteam` request was truncated, the response includes a `NextToken`. To retrieve the next set of labeling jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `CreationTime`

Required: No

SortOrder

The sort order for results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

WorkteamArn

The Amazon Resource Name (ARN) of the work team for which you want to see labeling jobs for.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:workteam/.*`

Required: Yes

Response Syntax

```
{
   "LabelingJobSummaryList": [
      {
         "CreationTime": number,
         "JobReferenceCode": "string",
         "LabelCounters": {
            "HumanLabeled": number,
            "PendingHuman": number,
            "Total": number
         },
         "LabelingJobName": "string",
         "NumberOfHumanWorkersPerDataObject": number,
         "WorkRequesterAccountId": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

LabelingJobSummaryList

An array of `LabelingJobSummary` objects, each describing a labeling job.

Type: Array of LabelingJobForWorkteamSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of labeling jobs, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListLineageGroups

Service: Amazon SageMaker Service

A list of lineage groups shared with your AWS account. For more information, see Cross-Account Lineage Tracking in the *Amazon SageMaker Developer Guide*.

Request Syntax

```
{
   "CreatedAfter": number,
   "CreatedBefore": number,
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreatedAfter

A timestamp to filter against lineage groups created after a certain point in time.

Type: Timestamp

Required: No

CreatedBefore

A timestamp to filter against lineage groups created before a certain point in time.

Type: Timestamp

Required: No

MaxResults

The maximum number of endpoints to return in the response. This value defaults to 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of algorithms, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The parameter by which to sort the results. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order for the results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "LineageGroupSummaries": [
      {
         "CreationTime": number,
         "DisplayName": "string",
```

```
            "LastModifiedTime": number,
            "LineageGroupArn": "string",
            "LineageGroupName": "string"
        }
    ],
    "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

LineageGroupSummaries

A list of lineage groups and their properties.

Type: Array of LineageGroupSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of algorithms, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListMlflowApps

Service: Amazon SageMaker Service

Lists all MLflow Apps

Request Syntax

```
{
   "AccountDefaultStatus": "string",
   "CreatedAfter": number,
   "CreatedBefore": number,
   "DefaultForDomainId": "string",
   "MaxResults": number,
   "MlflowVersion": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "Status": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AccountDefaultStatus

Filter for MLflow Apps with the specified `AccountDefaultStatus`.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

CreatedAfter

Use the `CreatedAfter` filter to only list MLflow Apps created after a specific date and time. Listed MLflow Apps are shown with a date and time such as `"2024-03-16T01:46:56+00:00"`. The `CreatedAfter` parameter takes in a Unix timestamp.

Type: Timestamp

Required: No

CreatedBefore

Use the `CreatedBefore` filter to only list MLflow Apps created before a specific date and time. Listed MLflow Apps are shown with a date and time such as `"2024-03-16T01:46:56+00:00"`. The `CreatedAfter` parameter takes in a Unix timestamp.

Type: Timestamp

Required: No

DefaultForDomainId

Filter for MLflow Apps with the specified default SageMaker Domain ID.

Type: String

Required: No

MaxResults

The maximum number of MLflow Apps to list.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

MlflowVersion

Filter for Mlflow Apps with the specified version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 16.

Pattern: `[0-9]*.[0-9]*.[0-9]*`

Required: No

NextToken

If the previous response was truncated, use this token in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Filter for MLflow Apps sorting by name, creation time, or creation status.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

Change the order of the listed MLflow Apps. By default, MLflow Apps are listed in `Descending` order by creation time. To change the list order, specify `SortOrder` to be `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Status

Filter for Mlflow apps with a specific creation status.

Type: String

Valid Values: `Creating | Created | CreateFailed | Updating | Updated | UpdateFailed | Deleting | DeleteFailed | Deleted`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "Summaries": [
      {
```

```
            "Arn": "string",
            "CreationTime": number,
            "LastModifiedTime": number,
            "MlflowVersion": "string",
            "Name": "string",
            "Status": "string"
        }
    ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Summaries

A list of MLflow Apps according to chosen filters.

Type: Array of MlflowAppSummary objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListMlflowTrackingServers

Service: Amazon SageMaker Service

Lists all MLflow Tracking Servers.

Request Syntax

```
{
   "CreatedAfter": number,
   "CreatedBefore": number,
   "MaxResults": number,
   "MlflowVersion": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "TrackingServerStatus": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreatedAfter

Use the `CreatedAfter` filter to only list tracking servers created after a specific date and time. Listed tracking servers are shown with a date and time such as `"2024-03-16T01:46:56+00:00"`. The `CreatedAfter` parameter takes in a Unix timestamp. To convert a date and time into a Unix timestamp, see EpochConverter.

Type: Timestamp

Required: No

CreatedBefore

Use the `CreatedBefore` filter to only list tracking servers created before a specific date and time. Listed tracking servers are shown with a date and time such as `"2024-03-16T01:46:56+00:00"`. The `CreatedBefore` parameter takes in a Unix timestamp. To convert a date and time into a Unix timestamp, see EpochConverter.

Type: Timestamp

Required: No

MaxResults

The maximum number of tracking servers to list.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

MlflowVersion

Filter for tracking servers using the specified MLflow version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 16.

Pattern: `[0-9]*.[0-9]*.[0-9]*`

Required: No

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Filter for trackings servers sorting by name, creation time, or creation status.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

Change the order of the listed tracking servers. By default, tracking servers are listed in `Descending` order by creation time. To change the list order, you can specify `SortOrder` to be `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

TrackingServerStatus

Filter for tracking servers with a specified creation status.

Type: String

Valid Values: `Creating | Created | CreateFailed | Updating | Updated | UpdateFailed | Deleting | DeleteFailed | Stopping | Stopped | StopFailed | Starting | Started | StartFailed | MaintenanceInProgress | MaintenanceComplete | MaintenanceFailed`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "TrackingServerSummaries": [
      {
         "CreationTime": number,
         "IsActive": "string",
         "LastModifiedTime": number,
         "MlflowVersion": "string",
         "TrackingServerArn": "string",
         "TrackingServerName": "string",
         "TrackingServerStatus": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

TrackingServerSummaries

A list of tracking servers according to chosen filters.

Type: Array of TrackingServerSummary objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListModelBiasJobDefinitions

Service: Amazon SageMaker Service

Lists model bias jobs definitions that satisfy various filters.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "EndpointName": "string",
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only model bias jobs created after a specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only model bias jobs created before a specified time.

Type: Timestamp

Required: No

EndpointName

Name of the endpoint to monitor for model bias.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MaxResults

The maximum number of model bias jobs to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

Filter for model bias jobs whose name contains a specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

The token returned if the response is truncated. To retrieve the next set of job executions, use it in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Whether to sort results by the `Name` or `CreationTime` field. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

Whether to sort the results in `Ascending` or `Descending` order. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "JobDefinitionSummaries": [
      {
         "CreationTime": number,
         "EndpointName": "string",
         "MonitoringJobDefinitionArn": "string",
         "MonitoringJobDefinitionName": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobDefinitionSummaries

A JSON array in which each element is a summary for a model bias jobs.

Type: Array of MonitoringJobDefinitionSummary objects

NextToken

The token returned if the response is truncated. To retrieve the next set of job executions, use it in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListModelCardExportJobs

Service: Amazon SageMaker Service

List the export jobs for the Amazon SageMaker Model Card.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "ModelCardExportJobNameContains": "string",
   "ModelCardName": "string",
   "ModelCardVersion": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Only list model card export jobs that were created after the time specified.

Type: Timestamp

Required: No

CreationTimeBefore

Only list model card export jobs that were created before the time specified.

Type: Timestamp

Required: No

MaxResults

The maximum number of model card export jobs to list.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

ModelCardExportJobNameContains

Only list model card export jobs with names that contain the specified string.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelCardName

List export jobs for the model card with the specified name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelCardVersion

List export jobs for the model card with the specified version.

Type: Integer

Required: No

NextToken

If the response to a previous `ListModelCardExportJobs` request was truncated, the response includes a `NextToken`. To retrieve the next set of model card export jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Sort model card export jobs by either name or creation time. Sorts by creation time by default.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

Sort model card export jobs by ascending or descending order.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

Only list model card export jobs with the specified status.

Type: String

Valid Values: `InProgress | Completed | Failed`

Required: No

Response Syntax

```
{
   "ModelCardExportJobSummaries": [
      {
         "CreatedAt": number,
         "LastModifiedAt": number,
         "ModelCardExportJobArn": "string",
         "ModelCardExportJobName": "string",
```

```
            "ModelCardName": "string",
            "ModelCardVersion": number,
            "Status": "string"
        }
    ],
    "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelCardExportJobSummaries

The summaries of the listed model card export jobs.

Type: Array of ModelCardExportJobSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of model card export jobs, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListModelCards

Service: Amazon SageMaker Service

List existing model cards.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "ModelCardStatus": "string",
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Only list model cards that were created after the time specified.

Type: Timestamp

Required: No

CreationTimeBefore

Only list model cards that were created before the time specified.

Type: Timestamp

Required: No

MaxResults

The maximum number of model cards to list.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

ModelCardStatus

Only list model cards with the specified approval status.

Type: String

Valid Values: `Draft | PendingReview | Approved | Archived`

Required: No

NameContains

Only list model cards with names that contain the specified string.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

NextToken

If the response to a previous `ListModelCards` request was truncated, the response includes a `NextToken`. To retrieve the next set of model cards, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Sort model cards by either name or creation time. Sorts by creation time by default.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

Sort model cards by ascending or descending order.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "ModelCardSummaries": [
      {
         "CreationTime": number,
         "LastModifiedTime": number,
         "ModelCardArn": "string",
         "ModelCardName": "string",
         "ModelCardStatus": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelCardSummaries

The summaries of the listed model cards.

Type: Array of ModelCardSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of model cards, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListModelCardVersions

Service: Amazon SageMaker Service

List existing versions of an Amazon SageMaker Model Card.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "ModelCardName": "string",
   "ModelCardStatus": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Only list model card versions that were created after the time specified.

Type: Timestamp

Required: No

CreationTimeBefore

Only list model card versions that were created before the time specified.

Type: Timestamp

Required: No

MaxResults

The maximum number of model card versions to list.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

ModelCardName

List model card versions for the model card with the specified name or Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:model-card/.*)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

ModelCardStatus

Only list model card versions with the specified approval status.

Type: String

Valid Values: `Draft | PendingReview | Approved | Archived`

Required: No

NextToken

If the response to a previous `ListModelCardVersions` request was truncated, the response includes a `NextToken`. To retrieve the next set of model card versions, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Sort listed model card versions by version. Sorts by version by default.

Type: String

Valid Values: `Version`

Required: No

SortOrder

Sort model card versions by ascending or descending order.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "ModelCardVersionSummaryList": [
      {
         "CreationTime": number,
         "LastModifiedTime": number,
         "ModelCardArn": "string",
         "ModelCardName": "string",
         "ModelCardStatus": "string",
         "ModelCardVersion": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelCardVersionSummaryList

The summaries of the listed versions of the model card.

Type: Array of ModelCardVersionSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of model card versions, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListModelExplainabilityJobDefinitions

Service: Amazon SageMaker Service

Lists model explainability job definitions that satisfy various filters.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "EndpointName": "string",
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only model explainability jobs created after a specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only model explainability jobs created before a specified time.

Type: Timestamp

Required: No

EndpointName

Name of the endpoint to monitor for model explainability.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MaxResults

The maximum number of jobs to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

Filter for model explainability jobs whose name contains a specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

The token returned if the response is truncated. To retrieve the next set of job executions, use it in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Whether to sort results by the `Name` or `CreationTime` field. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

Whether to sort the results in `Ascending` or `Descending` order. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "JobDefinitionSummaries": [
      {
         "CreationTime": number,
         "EndpointName": "string",
         "MonitoringJobDefinitionArn": "string",
         "MonitoringJobDefinitionName": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobDefinitionSummaries

A JSON array in which each element is a summary for a explainability bias jobs.

Type: Array of MonitoringJobDefinitionSummary objects

NextToken

The token returned if the response is truncated. To retrieve the next set of job executions, use it in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListModelMetadata

Service: Amazon SageMaker Service

Lists the domain, framework, task, and model name of standard machine learning models found in common model zoos.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string",
   "SearchExpression": {
      "Filters": [
         {
            "Name": "string",
            "Value": "string"
         }
      ]
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

The maximum number of models to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the response to a previous `ListModelMetadataResponse` request was truncated, the response includes a NextToken. To retrieve the next set of model metadata, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SearchExpression

One or more filters that searches for the specified resource or resources in a search. All resource objects that satisfy the expression's condition are included in the search results. Specify the Framework, FrameworkVersion, Domain or Task to filter supported. Filter names and values are case-sensitive.

Type: ModelMetadataSearchExpression object

Required: No

Response Syntax

```
{
   "ModelMetadataSummaries": [
      {
         "Domain": "string",
         "Framework": "string",
         "FrameworkVersion": "string",
         "Model": "string",
         "Task": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelMetadataSummaries

A structure that holds model metadata.

Type: Array of ModelMetadataSummary objects

NextToken

A token for getting the next set of recommendations, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListModelPackageGroups

Service: Amazon SageMaker Service

Gets a list of the model groups in your AWS account.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "CrossAccountFilterOption": "string",
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only model groups created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only model groups created before the specified time.

Type: Timestamp

Required: No

CrossAccountFilterOption

A filter that returns either model groups shared with you or model groups in your own account. When the value is `CrossAccount`, the results show the resources made discoverable to you

from other accounts. When the value is `SameAccount` or `null`, the results show resources from your account. The default is `SameAccount`.

Type: String

Valid Values: `SameAccount | CrossAccount`

Required: No

MaxResults

The maximum number of results to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the model group name. This filter returns only model groups whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the result of the previous `ListModelPackageGroups` request was truncated, the response includes a `NextToken`. To retrieve the next set of model groups, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order for results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "ModelPackageGroupSummaryList": [
      {
         "CreationTime": number,
         "ManagedConfiguration": {
            "ManagedStorageType": "string"
         },
         "ModelPackageGroupArn": "string",
         "ModelPackageGroupDescription": "string",
         "ModelPackageGroupName": "string",
         "ModelPackageGroupStatus": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelPackageGroupSummaryList

A list of summaries of the model groups in your AWS account.

Type: Array of ModelPackageGroupSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of model groups, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListModelPackages

Service: Amazon SageMaker Service

Lists the model packages that have been created.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "ModelApprovalStatus": "string",
   "ModelPackageGroupName": "string",
   "ModelPackageType": "string",
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only model packages created after the specified time (timestamp).

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only model packages created before the specified time (timestamp).

Type: Timestamp

Required: No

MaxResults

The maximum number of model packages to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

ModelApprovalStatus

A filter that returns only the model packages with the specified approval status.

Type: String

Valid Values: `Approved | Rejected | PendingManualApproval`

Required: No

ModelPackageGroupName

A filter that returns only model versions that belong to the specified model group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 170.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*\/)?([a-zA-Z0-9]([a-zA-Z0-9-]){0,62})(?<!-)`

Required: No

ModelPackageType

A filter that returns only the model packages of the specified type. This can be one of the following values.

- `UNVERSIONED` - List only unversioined models. This is the default value if no `ModelPackageType` is specified.
- `VERSIONED` - List only versioned models.
- `BOTH` - List both versioned and unversioned models.

Type: String

Valid Values: `Versioned | Unversioned | Both`

Required: No

NameContains

A string in the model package name. This filter returns only model packages whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the response to a previous `ListModelPackages` request was truncated, the response includes a `NextToken`. To retrieve the next set of model packages, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The parameter by which to sort the results. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order for the results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "ModelPackageSummaryList": [
      {
         "CreationTime": number,
         "ModelApprovalStatus": "string",
         "ModelLifeCycle": {
            "Stage": "string",
            "StageDescription": "string",
            "StageStatus": "string"
         },
         "ModelPackageArn": "string",
         "ModelPackageDescription": "string",
         "ModelPackageGroupName": "string",
         "ModelPackageName": "string",
         "ModelPackageRegistrationType": "string",
         "ModelPackageStatus": "string",
         "ModelPackageVersion": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelPackageSummaryList

An array of `ModelPackageSummary` objects, each of which lists a model package.

Type: Array of ModelPackageSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of model packages, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListModelQualityJobDefinitions

Service: Amazon SageMaker Service

Gets a list of model quality monitoring job definitions in your account.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "EndpointName": "string",
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only model quality monitoring job definitions created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only model quality monitoring job definitions created before the specified time.

Type: Timestamp

Required: No

EndpointName

A filter that returns only model quality monitoring job definitions that are associated with the specified endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MaxResults

The maximum number of results to return in a call to `ListModelQualityJobDefinitions`.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the transform job name. This filter returns only model quality monitoring job definitions whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the result of the previous `ListModelQualityJobDefinitions` request was truncated, the response includes a `NextToken`. To retrieve the next set of model quality monitoring job definitions, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

Whether to sort the results in `Ascending` or `Descending` order. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "JobDefinitionSummaries": [
      {
         "CreationTime": number,
         "EndpointName": "string",
         "MonitoringJobDefinitionArn": "string",
         "MonitoringJobDefinitionName": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

JobDefinitionSummaries

A list of summaries of model quality monitoring job definitions.

Type: Array of MonitoringJobDefinitionSummary objects

NextToken

If the response is truncated, Amazon SageMaker AI returns this token. To retrieve the next set of model quality monitoring job definitions, use it in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListModels

Service: Amazon SageMaker Service

Lists models created with the `CreateModel` API.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only models with a creation time greater than or equal to the specified time (timestamp).

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only models created before the specified time (timestamp).

Type: Timestamp

Required: No

MaxResults

The maximum number of models to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the model name. This filter returns only models whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9-]+`

Required: No

NextToken

If the response to a previous `ListModels` request was truncated, the response includes a `NextToken`. To retrieve the next set of models, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Sorts the list of results. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order for results. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "Models": [
      {
         "CreationTime": number,
         "ModelArn": "string",
         "ModelName": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Models

An array of `ModelSummary` objects, each of which lists a model.

Type: Array of ModelSummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of models, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListMonitoringAlertHistory

Service: Amazon SageMaker Service

Gets a list of past alerts in a model monitoring schedule.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "MonitoringAlertName": "string",
   "MonitoringScheduleName": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only alerts created on or after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only alerts created on or before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of results to display. The default is 100.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

MonitoringAlertName

The name of a monitoring alert.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MonitoringScheduleName

The name of a monitoring schedule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

NextToken

If the result of the previous `ListMonitoringAlertHistory` request was truncated, the response includes a `NextToken`. To retrieve the next set of alerts in the history, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field used to sort results. The default is `CreationTime`.

Type: String

Valid Values: `CreationTime | Status`

Required: No

SortOrder

The sort order, whether `Ascending` or `Descending`, of the alert history. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that retrieves only alerts with a specific status.

Type: String

Valid Values: `InAlert | OK`

Required: No

Response Syntax

```
{
   "MonitoringAlertHistory": [
      {
         "AlertStatus": "string",
         "CreationTime": number,
         "MonitoringAlertName": "string",
         "MonitoringScheduleName": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

MonitoringAlertHistory

An alert history for a model monitoring schedule.

Type: Array of MonitoringAlertHistorySummary objects

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of alerts, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListMonitoringAlerts

Service: Amazon SageMaker Service

Gets the alerts for a single monitoring schedule.

Request Syntax

```
{
   "MaxResults": number,
   "MonitoringScheduleName": "string",
   "NextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

The maximum number of results to display. The default is 100.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

MonitoringScheduleName

The name of a monitoring schedule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

NextToken

If the result of the previous `ListMonitoringAlerts` request was truncated, the response includes a `NextToken`. To retrieve the next set of alerts in the history, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "MonitoringAlertSummaries": [
      {
         "Actions": {
            "ModelDashboardIndicator": {
               "Enabled": boolean
            }
         },
         "AlertStatus": "string",
         "CreationTime": number,
         "DatapointsToAlert": number,
         "EvaluationPeriod": number,
         "LastModifiedTime": number,
         "MonitoringAlertName": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

MonitoringAlertSummaries

A JSON array where each element is a summary for a monitoring alert.

Type: Array of MonitoringAlertSummary objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of alerts, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListMonitoringExecutions

Service: Amazon SageMaker Service

Returns list of all monitoring job executions.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "EndpointName": "string",
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "MonitoringJobDefinitionName": "string",
   "MonitoringScheduleName": "string",
   "MonitoringTypeEquals": "string",
   "NextToken": "string",
   "ScheduledTimeAfter": number,
   "ScheduledTimeBefore": number,
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only jobs created after a specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only jobs created before a specified time.

Type: Timestamp

Required: No

EndpointName

Name of a specific endpoint to fetch jobs for.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

LastModifiedTimeAfter

A filter that returns only jobs modified before a specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only jobs modified after a specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of jobs to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

MonitoringJobDefinitionName

Gets a list of the monitoring job runs of the specified monitoring job definitions.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MonitoringScheduleName

Name of a specific schedule to fetch jobs for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MonitoringTypeEquals

A filter that returns only the monitoring job runs of the specified monitoring type.

Type: String

Valid Values: `DataQuality | ModelQuality | ModelBias | ModelExplainability`

Required: No

NextToken

The token returned if the response is truncated. To retrieve the next set of job executions, use it in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

ScheduledTimeAfter

Filter for jobs scheduled after a specified time.

Type: Timestamp

Required: No

ScheduledTimeBefore

Filter for jobs scheduled before a specified time.

Type: Timestamp

Required: No

SortBy

Whether to sort the results by the `Status`, `CreationTime`, or `ScheduledTime` field. The default is `CreationTime`.

Type: String

Valid Values: `CreationTime | ScheduledTime | Status`

Required: No

SortOrder

Whether to sort the results in `Ascending` or `Descending` order. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that retrieves only jobs with a specific status.

Type: String

Valid Values: `Pending | Completed | CompletedWithViolations | InProgress | Failed | Stopping | Stopped`

Required: No

Response Syntax

```
{
```

```
   "MonitoringExecutionSummaries": [
      {
         "CreationTime": number,
         "EndpointName": "string",
         "FailureReason": "string",
         "LastModifiedTime": number,
         "MonitoringExecutionStatus": "string",
         "MonitoringJobDefinitionName": "string",
         "MonitoringScheduleName": "string",
         "MonitoringType": "string",
         "ProcessingJobArn": "string",
         "ScheduledTime": number
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

MonitoringExecutionSummaries

A JSON array in which each element is a summary for a monitoring execution.

Type: Array of MonitoringExecutionSummary objects

NextToken

The token returned if the response is truncated. To retrieve the next set of job executions, use it in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListMonitoringSchedules

Service: Amazon SageMaker Service

Returns list of all monitoring schedules.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "EndpointName": "string",
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "MonitoringJobDefinitionName": "string",
   "MonitoringTypeEquals": "string",
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only monitoring schedules created after a specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only monitoring schedules created before a specified time.

Type: Timestamp

Required: No

EndpointName

Name of a specific endpoint to fetch schedules for.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

LastModifiedTimeAfter

A filter that returns only monitoring schedules modified after a specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only monitoring schedules modified before a specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of jobs to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

MonitoringJobDefinitionName

Gets a list of the monitoring schedules for the specified monitoring job definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MonitoringTypeEquals

A filter that returns only the monitoring schedules for the specified monitoring type.

Type: String

Valid Values: `DataQuality | ModelQuality | ModelBias | ModelExplainability`

Required: No

NameContains

Filter for monitoring schedules whose name contains a specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

The token returned if the response is truncated. To retrieve the next set of job executions, use it in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Whether to sort the results by the `Status`, `CreationTime`, or `ScheduledTime` field. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

Whether to sort the results in `Ascending` or `Descending` order. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that returns only monitoring schedules modified before a specified time.

Type: String

Valid Values: `Pending | Failed | Scheduled | Stopped`

Required: No

Response Syntax

```
{
   "MonitoringScheduleSummaries": [
      {
         "CreationTime": number,
         "EndpointName": "string",
         "LastModifiedTime": number,
         "MonitoringJobDefinitionName": "string",
         "MonitoringScheduleArn": "string",
         "MonitoringScheduleName": "string",
         "MonitoringScheduleStatus": "string",
         "MonitoringType": "string"
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

MonitoringScheduleSummaries

A JSON array in which each element is a summary for a monitoring schedule.

Type: Array of MonitoringScheduleSummary objects

NextToken

The token returned if the response is truncated. To retrieve the next set of job executions, use it in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListNotebookInstanceLifecycleConfigs

Service: Amazon SageMaker Service

Lists notebook instance lifestyle configurations created with the CreateNotebookInstanceLifecycleConfig API.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only lifecycle configurations that were created after the specified time (timestamp).

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only lifecycle configurations that were created before the specified time (timestamp).

Type: Timestamp

Required: No

LastModifiedTimeAfter

A filter that returns only lifecycle configurations that were modified after the specified time (timestamp).

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only lifecycle configurations that were modified before the specified time (timestamp).

Type: Timestamp

Required: No

MaxResults

The maximum number of lifecycle configurations to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the lifecycle configuration name. This filter returns only lifecycle configurations whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9-]+`

Required: No

NextToken

If the result of a `ListNotebookInstanceLifecycleConfigs` request was truncated, the response includes a `NextToken`. To get the next set of lifecycle configurations, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Sorts the list of results. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime | LastModifiedTime`

Required: No

SortOrder

The sort order for results.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "NotebookInstanceLifecycleConfigs": [
      {
         "CreationTime": number,
         "LastModifiedTime": number,
         "NotebookInstanceLifecycleConfigArn": "string",
         "NotebookInstanceLifecycleConfigName": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the response is truncated, SageMaker AI returns this token. To get the next set of lifecycle configurations, use it in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

NotebookInstanceLifecycleConfigs

An array of `NotebookInstanceLifecycleConfiguration` objects, each listing a lifecycle configuration.

Type: Array of NotebookInstanceLifecycleConfigSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListNotebookInstances

Service: Amazon SageMaker Service

Returns a list of the SageMaker AI notebook instances in the requester's account in an AWS Region.

Request Syntax

```
{
   "AdditionalCodeRepositoryEquals": "string",
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "DefaultCodeRepositoryContains": "string",
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "NotebookInstanceLifecycleConfigNameContains": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AdditionalCodeRepositoryEquals

A filter that returns only notebook instances with associated with the specified git repository.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `https://([^/]+)/?(.*)$|^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

CreationTimeAfter

A filter that returns only notebook instances that were created after the specified time (timestamp).

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only notebook instances that were created before the specified time (timestamp).

Type: Timestamp

Required: No

DefaultCodeRepositoryContains

A string in the name or URL of a Git repository associated with this notebook instance. This filter returns only notebook instances associated with a git repository with a name that contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[a-zA-Z0-9-]+`

Required: No

LastModifiedTimeAfter

A filter that returns only notebook instances that were modified after the specified time (timestamp).

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only notebook instances that were modified before the specified time (timestamp).

Type: Timestamp

Required: No

MaxResults

The maximum number of notebook instances to return.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the notebook instances' name. This filter returns only notebook instances whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9-]+`

Required: No

NextToken

If the previous call to the `ListNotebookInstances` is truncated, the response includes a `NextToken`. You can use this token in your subsequent `ListNotebookInstances` request to fetch the next set of notebook instances.

> **Note**
>
> You might specify a filter or a sort order in your request. When response is truncated, you must use the same values for the filer and sort order in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

NotebookInstanceLifecycleConfigNameContains

A string in the name of a notebook instances lifecycle configuration associated with this notebook instance. This filter returns only notebook instances associated with a lifecycle configuration with a name that contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

SortBy

The field to sort results by. The default is `Name`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that returns only notebook instances with the specified status.

Type: String

Valid Values: `Pending | InService | Stopping | Stopped | Failed | Deleting | Updating`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "NotebookInstances": [
      {
         "AdditionalCodeRepositories": [ "string" ],
         "CreationTime": number,
         "DefaultCodeRepository": "string",
         "InstanceType": "string",
         "LastModifiedTime": number,
         "NotebookInstanceArn": "string",
         "NotebookInstanceLifecycleConfigName": "string",
         "NotebookInstanceName": "string",
         "NotebookInstanceStatus": "string",
         "Url": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the response to the previous `ListNotebookInstances` request was truncated, SageMaker AI returns this token. To retrieve the next set of notebook instances, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

NotebookInstances

An array of `NotebookInstanceSummary` objects, one for each notebook instance.

Type: Array of NotebookInstanceSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListOptimizationJobs

Service: Amazon SageMaker Service

Lists the optimization jobs in your account and their properties.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "OptimizationContains": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Filters the results to only those optimization jobs that were created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

Filters the results to only those optimization jobs that were created before the specified time.

Type: Timestamp

Required: No

LastModifiedTimeAfter

Filters the results to only those optimization jobs that were updated after the specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

Filters the results to only those optimization jobs that were updated before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of optimization jobs to return in the response. The default is 50.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

Filters the results to only those optimization jobs with a name that contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

A token that you use to get the next set of results following a truncated response. If the response to the previous request was truncated, that response provides the value for this token.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

OptimizationContains

Filters the results to only those optimization jobs that apply the specified optimization techniques. You can specify either `Quantization` or `Compilation`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

SortBy

The field by which to sort the optimization jobs in the response. The default is `CreationTime`

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results. The default is `Ascending`

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

Filters the results to only those optimization jobs with the specified status.

Type: String

Valid Values: `INPROGRESS | COMPLETED | FAILED | STARTING | STOPPING | STOPPED`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "OptimizationJobSummaries": [
      {
         "CreationTime": number,
         "DeploymentInstanceType": "string",
         "LastModifiedTime": number,
         "MaxInstanceCount": number,
         "OptimizationEndTime": number,
         "OptimizationJobArn": "string",
         "OptimizationJobName": "string",
         "OptimizationJobStatus": "string",
         "OptimizationStartTime": number,
         "OptimizationTypes": [ "string" ]
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

The token to use in a subsequent request to get the next set of results following a truncated response.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

OptimizationJobSummaries

A list of optimization jobs and their properties that matches any of the filters you specified in the request.

Type: Array of OptimizationJobSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListPartnerApps

Service: Amazon SageMaker Service

Lists all of the SageMaker Partner AI Apps in an account.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

This parameter defines the maximum number of results that can be returned in a single response. The `MaxResults` parameter is an upper bound, not a target. If there are more results available than the value specified, a `NextToken` is provided in the response. The `NextToken` indicates that the user should get the next set of results by providing this token as a part of a subsequent call. The default value for `MaxResults` is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "Summaries": [
      {
         "Arn": "string",
         "CreationTime": number,
         "Name": "string",
         "Status": "string",
         "Type": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Summaries

The information related to each of the SageMaker Partner AI Apps in an account.

Type: Array of PartnerAppSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListPipelineExecutions

Service: Amazon SageMaker Service

Gets a list of the pipeline executions.

Request Syntax

```
{
   "CreatedAfter": number,
   "CreatedBefore": number,
   "MaxResults": number,
   "NextToken": "string",
   "PipelineName": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreatedAfter

A filter that returns the pipeline executions that were created after a specified time.

Type: Timestamp

Required: No

CreatedBefore

A filter that returns the pipeline executions that were created before a specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of pipeline executions to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the result of the previous `ListPipelineExecutions` request was truncated, the response includes a `NextToken`. To retrieve the next set of pipeline executions, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

PipelineName

The name or Amazon Resource Name (ARN) of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,255})`

Required: Yes

SortBy

The field by which to sort results. The default is `CreatedTime`.

Type: String

Valid Values: `CreationTime | PipelineExecutionArn`

Required: No

SortOrder

The sort order for results.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "PipelineExecutionSummaries": [
      {
         "PipelineExecutionArn": "string",
         "PipelineExecutionDescription": "string",
         "PipelineExecutionDisplayName": "string",
         "PipelineExecutionFailureReason": "string",
         "PipelineExecutionStatus": "string",
         "StartTime": number
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the result of the previous `ListPipelineExecutions` request was truncated, the response includes a `NextToken`. To retrieve the next set of pipeline executions, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

PipelineExecutionSummaries

Contains a sorted list of pipeline execution summary objects matching the specified filters. Each run summary includes the Amazon Resource Name (ARN) of the pipeline execution, the run date, and the status. This list can be empty.

Type: Array of PipelineExecutionSummary objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListPipelineExecutionSteps

Service: Amazon SageMaker Service

Gets a list of `PipeLineExecutionStep` objects.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string",
   "PipelineExecutionArn": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

The maximum number of pipeline execution steps to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the result of the previous `ListPipelineExecutionSteps` request was truncated, the response includes a `NextToken`. To retrieve the next set of pipeline execution steps, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: No

SortOrder

The field by which to sort results. The default is `CreatedTime`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "PipelineExecutionSteps": [
      {
         "AttemptCount": number,
         "CacheHitResult": {
            "SourcePipelineExecutionArn": "string"
         },
         "EndTime": number,
         "FailureReason": "string",
         "Metadata": {
            "AutoMLJob": {
               "Arn": "string"
            },
            "BedrockCustomModel": {
               "Arn": "string"
            },
            "BedrockCustomModelDeployment": {
               "Arn": "string"
```

```
            },
            "BedrockModelImport": {
               "Arn": "string"
            },
            "BedrockProvisionedModelThroughput": {
               "Arn": "string"
            },
            "Callback": {
               "CallbackToken": "string",
               "OutputParameters": [
                  {
                     "Name": "string",
                     "Value": "string"
                  }
               ],
               "SqsQueueUrl": "string"
            },
            "ClarifyCheck": {
               "BaselineUsedForDriftCheckConstraints": "string",
               "CalculatedBaselineConstraints": "string",
               "CheckJobArn": "string",
               "CheckType": "string",
               "ModelPackageGroupName": "string",
               "RegisterNewBaseline": boolean,
               "SkipCheck": boolean,
               "ViolationReport": "string"
            },
            "Condition": {
               "Outcome": "string"
            },
            "EMR": {
               "ClusterId": "string",
               "LogFilePath": "string",
               "StepId": "string",
               "StepName": "string"
            },
            "Endpoint": {
               "Arn": "string"
            },
            "EndpointConfig": {
               "Arn": "string"
            },
            "Fail": {
               "ErrorMessage": "string"
```

```
            },
            "InferenceComponent": {
               "Arn": "string"
            },
            "Job": {
               "Arn": "string"
            },
            "Lambda": {
               "Arn": "string",
               "OutputParameters": [
                  {
                     "Name": "string",
                     "Value": "string"
                  }
               ]
            },
            "Lineage": {
               "ActionArns": {
                  "string" : "string"
               },
               "ArtifactArns": {
                  "string" : "string"
               },
               "Associations": [
                  {
                     "DestinationArn": "string",
                     "SourceArn": "string"
                  }
               ],
               "ContextArns": {
                  "string" : "string"
               }
            },
            "Model": {
               "Arn": "string"
            },
            "ProcessingJob": {
               "Arn": "string"
            },
            "QualityCheck": {
               "BaselineUsedForDriftCheckConstraints": "string",
               "BaselineUsedForDriftCheckStatistics": "string",
               "CalculatedBaselineConstraints": "string",
               "CalculatedBaselineStatistics": "string",
```

```
                "CheckJobArn": "string",
                "CheckType": "string",
                "ModelPackageGroupName": "string",
                "RegisterNewBaseline": boolean,
                "SkipCheck": boolean,
                "ViolationReport": "string"
            },
            "RegisterModel": {
                "Arn": "string"
            },
            "TrainingJob": {
                "Arn": "string"
            },
            "TransformJob": {
                "Arn": "string"
            },
            "TuningJob": {
                "Arn": "string"
            }
        },
        "SelectiveExecutionResult": {
            "SourcePipelineExecutionArn": "string"
        },
        "StartTime": number,
        "StepDescription": "string",
        "StepDisplayName": "string",
        "StepName": "string",
        "StepStatus": "string"
    }
  ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the result of the previous `ListPipelineExecutionSteps` request was truncated, the response includes a `NextToken`. To retrieve the next set of pipeline execution steps, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

PipelineExecutionSteps

A list of `PipeLineExecutionStep` objects. Each `PipeLineExecutionStep` consists of StepName, StartTime, EndTime, StepStatus, and Metadata. Metadata is an object with properties for each job that contains relevant information about the job created by the step.

Type: Array of PipelineExecutionStep objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

ListPipelineParametersForExecution

Service: Amazon SageMaker Service

Gets a list of parameters for a pipeline execution.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string",
   "PipelineExecutionArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

The maximum number of parameters to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the result of the previous `ListPipelineParametersForExecution` request was truncated, the response includes a `NextToken`. To retrieve the next set of parameters, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: Yes

Response Syntax

```
{
   "NextToken": "string",
   "PipelineParameters": [
      {
         "Name": "string",
         "Value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the result of the previous `ListPipelineParametersForExecution` request was truncated, the response includes a `NextToken`. To retrieve the next set of parameters, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

PipelineParameters

Contains a list of pipeline parameters. This list can be empty.

Type: Array of Parameter objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListPipelines

Service: Amazon SageMaker Service

Gets a list of pipelines.

Request Syntax

```
{
   "CreatedAfter": number,
   "CreatedBefore": number,
   "MaxResults": number,
   "NextToken": "string",
   "PipelineNamePrefix": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreatedAfter

A filter that returns the pipelines that were created after a specified time.

Type: Timestamp

Required: No

CreatedBefore

A filter that returns the pipelines that were created before a specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of pipelines to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the result of the previous `ListPipelines` request was truncated, the response includes a `NextToken`. To retrieve the next set of pipelines, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

PipelineNamePrefix

The prefix of the pipeline name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: No

SortBy

The field by which to sort results. The default is `CreatedTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order for results.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "PipelineSummaries": [
      {
         "CreationTime": number,
         "LastExecutionTime": number,
         "LastModifiedTime": number,
         "PipelineArn": "string",
         "PipelineDescription": "string",
         "PipelineDisplayName": "string",
         "PipelineName": "string",
         "RoleArn": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the result of the previous `ListPipelines` request was truncated, the response includes a `NextToken`. To retrieve the next set of pipelines, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

PipelineSummaries

Contains a sorted list of `PipelineSummary` objects matching the specified filters. Each `PipelineSummary` consists of PipelineArn, PipelineName, ExperimentName, PipelineDescription, CreationTime, LastModifiedTime, LastRunTime, and RoleArn. This list can be empty.

Type: Array of PipelineSummary objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListPipelineVersions

Service: Amazon SageMaker Service

Gets a list of all versions of the pipeline.

Request Syntax

```
{
   "CreatedAfter": number,
   "CreatedBefore": number,
   "MaxResults": number,
   "NextToken": "string",
   "PipelineName": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreatedAfter

A filter that returns the pipeline versions that were created after a specified time.

Type: Timestamp

Required: No

CreatedBefore

A filter that returns the pipeline versions that were created before a specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of pipeline versions to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the result of the previous `ListPipelineVersions` request was truncated, the response includes a `NextToken`. To retrieve the next set of pipeline versions, use this token in your next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

PipelineName

The Amazon Resource Name (ARN) of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,255})`

Required: Yes

SortOrder

The sort order for the results.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "NextToken": "string",
```

```
   "PipelineVersionSummaries": [
      {
         "CreationTime": number,
         "LastExecutionPipelineExecutionArn": "string",
         "PipelineArn": "string",
         "PipelineVersionDescription": "string",
         "PipelineVersionDisplayName": "string",
         "PipelineVersionId": number
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the result of the previous `ListPipelineVersions` request was truncated, the response includes a `NextToken`. To retrieve the next set of pipeline versions, use this token in your next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

PipelineVersionSummaries

Contains a sorted list of pipeline version summary objects matching the specified filters. Each version summary includes the pipeline version ID, the creation date, and the last pipeline execution created from that version. This list can be empty.

Type: Array of PipelineVersionSummary objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListProcessingJobs

Service: Amazon SageMaker Service

Lists processing jobs that satisfy various filters.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only processing jobs created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only processing jobs created after the specified time.

Type: Timestamp

Required: No

LastModifiedTimeAfter

A filter that returns only processing jobs modified after the specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only processing jobs modified before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of processing jobs to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the processing job name. This filter returns only processing jobs whose name contains the specified string.

Type: String

Required: No

NextToken

If the result of the previous `ListProcessingJobs` request was truncated, the response includes a `NextToken`. To retrieve the next set of processing jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that retrieves only processing jobs with a specific status.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "ProcessingJobSummaries": [
      {
         "CreationTime": number,
         "ExitMessage": "string",
         "FailureReason": "string",
         "LastModifiedTime": number,
         "ProcessingEndTime": number,
         "ProcessingJobArn": "string",
         "ProcessingJobName": "string",
         "ProcessingJobStatus": "string"
```

```
        }
    ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the response is truncated, Amazon SageMaker returns this token. To retrieve the next set of processing jobs, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

ProcessingJobSummaries

An array of `ProcessingJobSummary` objects, each listing a processing job.

Type: Array of ProcessingJobSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListProjects

Service: Amazon SageMaker Service

Gets a list of the projects in an AWS account.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns the projects that were created after a specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns the projects that were created before a specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of projects to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A filter that returns the projects whose name contains a specified string.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: No

NextToken

If the result of the previous `ListProjects` request was truncated, the response includes a `NextToken`. To retrieve the next set of projects, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field by which to sort results. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order for results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "ProjectSummaryList": [
      {
         "CreationTime": number,
         "ProjectArn": "string",
         "ProjectDescription": "string",
         "ProjectId": "string",
         "ProjectName": "string",
         "ProjectStatus": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the result of the previous `ListCompilationJobs` request was truncated, the response includes a `NextToken`. To retrieve the next set of model compilation jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

ProjectSummaryList

A list of summaries of projects.

Type: Array of ProjectSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListResourceCatalogs

Service: Amazon SageMaker Service

Lists Amazon SageMaker Catalogs based on given filters and orders. The maximum number of `ResourceCatalogs` viewable is 1000.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

Use this parameter to search for `ResourceCatalogs` created after a specific date and time.

Type: Timestamp

Required: No

CreationTimeBefore

Use this parameter to search for `ResourceCatalogs` created before a specific date and time.

Type: Timestamp

Required: No

MaxResults

The maximum number of results returned by `ListResourceCatalogs`.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string that partially matches one or more `ResourceCatalogs` names. Filters `ResourceCatalog` by name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Required: No

NextToken

A token to resume pagination of `ListResourceCatalogs` results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The value on which the resource catalog list is sorted.

Type: String

Valid Values: `CreationTime`

Required: No

SortOrder

The order in which the resource catalogs are listed.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "ResourceCatalogs": [
      {
         "CreationTime": number,
         "Description": "string",
         "ResourceCatalogArn": "string",
         "ResourceCatalogName": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

A token to resume pagination of `ListResourceCatalogs` results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

ResourceCatalogs

A list of the requested `ResourceCatalogs`.

Type: Array of ResourceCatalog objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListSpaces

Service: Amazon SageMaker Service

Lists spaces.

Request Syntax

```
{
   "DomainIdEquals": "string",
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "SpaceNameContains": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainIdEquals

A parameter to search for the domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: No

MaxResults

This parameter defines the maximum number of results that can be return in a single response. The `MaxResults` parameter is an upper bound, not a target. If there are more results available than the value specified, a `NextToken` is provided in the response. The `NextToken` indicates that the user should get the next set of results by providing this token as a part of a subsequent call. The default value for `MaxResults` is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The parameter by which to sort the results. The default is `CreationTime`.

Type: String

Valid Values: `CreationTime | LastModifiedTime`

Required: No

SortOrder

The sort order for the results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

SpaceNameContains

A parameter by which to filter the results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "Spaces": [
      {
         "CreationTime": number,
         "DomainId": "string",
         "LastModifiedTime": number,
         "OwnershipSettingsSummary": {
            "OwnerUserProfileName": "string"
         },
         "SpaceDisplayName": "string",
         "SpaceName": "string",
         "SpaceSettingsSummary": {
            "AppType": "string",
            "RemoteAccess": "string",
            "SpaceStorageSettings": {
               "EbsStorageSettings": {
                  "EbsVolumeSizeInGb": number
               }
            }
         },
         "SpaceSharingSettingsSummary": {
            "SharingType": "string"
         },
         "Status": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Spaces

The list of spaces.

Type: Array of SpaceDetails objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListStageDevices

Service: Amazon SageMaker Service

Lists devices allocated to the stage, containing detailed device information and deployment status.

Request Syntax

```
{
   "EdgeDeploymentPlanName": "string",
   "ExcludeDevicesDeployedInOtherStage": boolean,
   "MaxResults": number,
   "NextToken": "string",
   "StageName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EdgeDeploymentPlanName

The name of the edge deployment plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ExcludeDevicesDeployedInOtherStage

Toggle for excluding devices deployed in other stages.

Type: Boolean

Required: No

MaxResults

The maximum number of requests to select.

Type: Integer

Valid Range: Maximum value of 100.

Required: No

NextToken

The response from the last list when returning a list large enough to neeed tokening.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

StageName

The name of the stage in the deployment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "DeviceDeploymentSummaries": [
      {
         "DeployedStageName": "string",
         "DeploymentStartTime": number,
         "Description": "string",
         "DeviceArn": "string",
         "DeviceDeploymentStatus": "string",
         "DeviceDeploymentStatusMessage": "string",
         "DeviceFleetName": "string",
         "DeviceName": "string",
         "EdgeDeploymentPlanArn": "string",
```

```
            "EdgeDeploymentPlanName": "string",
            "StageName": "string"
        }
    ],
    "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

DeviceDeploymentSummaries

List of summaries of devices allocated to the stage.

Type: Array of DeviceDeploymentSummary objects

NextToken

The token to use when calling the next page of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListStudioLifecycleConfigs

Service: Amazon SageMaker Service

Lists the Amazon SageMaker AI Studio Lifecycle Configurations in your AWS Account.

Request Syntax

```
{
   "AppTypeEquals": "string",
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "MaxResults": number,
   "ModifiedTimeAfter": number,
   "ModifiedTimeBefore": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AppTypeEquals

A parameter to search for the App Type to which the Lifecycle Configuration is attached.

Type: String

Valid Values: `JupyterServer | KernelGateway | CodeEditor | JupyterLab`

Required: No

CreationTimeAfter

A filter that returns only Lifecycle Configurations created on or after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only Lifecycle Configurations created on or before the specified time.

Type: Timestamp

Required: No

MaxResults

The total number of items to return in the response. If the total number of items available is more than the value specified, a `NextToken` is provided in the response. To resume pagination, provide the `NextToken` value in the as part of a subsequent call. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

ModifiedTimeAfter

A filter that returns only Lifecycle Configurations modified after the specified time.

Type: Timestamp

Required: No

ModifiedTimeBefore

A filter that returns only Lifecycle Configurations modified before the specified time.

Type: Timestamp

Required: No

NameContains

A string in the Lifecycle Configuration name. This filter returns only Lifecycle Configurations whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

NextToken

If the previous call to ListStudioLifecycleConfigs didn't return the full set of Lifecycle Configurations, the call returns a token for getting the next set of Lifecycle Configurations.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The property used to sort results. The default value is CreationTime.

Type: String

Valid Values: `CreationTime | LastModifiedTime | Name`

Required: No

SortOrder

The sort order. The default value is Descending.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "StudioLifecycleConfigs": [
      {
         "CreationTime": number,
         "LastModifiedTime": number,
         "StudioLifecycleConfigAppType": "string",
         "StudioLifecycleConfigArn": "string",
```

```
            "StudioLifecycleConfigName": "string"
        }
    ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

StudioLifecycleConfigs

A list of Lifecycle Configurations and their properties.

Type: Array of StudioLifecycleConfigDetails objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListSubscribedWorkteams

Service: Amazon SageMaker Service

Gets a list of the work teams that you are subscribed to in the AWS Marketplace. The list may be empty if no work team satisfies the filter specified in the `NameContains` parameter.

Request Syntax

```
{
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

The maximum number of work teams to return in each page of the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the work team name. This filter returns only work teams whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

NextToken

If the result of the previous `ListSubscribedWorkteams` request was truncated, the response includes a `NextToken`. To retrieve the next set of labeling jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "SubscribedWorkteams": [
      {
         "ListingId": "string",
         "MarketplaceDescription": "string",
         "MarketplaceTitle": "string",
         "SellerName": "string",
         "WorkteamArn": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the response is truncated, Amazon SageMaker returns this token. To retrieve the next set of work teams, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

SubscribedWorkteams

An array of `Workteam` objects, each describing a work team.

Type: Array of SubscribedWorkteam objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListTags

Service: Amazon SageMaker Service

Returns the tags for the specified SageMaker resource.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string",
   "ResourceArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

Maximum number of tags to return.

Type: Integer

Valid Range: Minimum value of 50.

Required: No

NextToken

If the response to the previous `ListTags` request is truncated, SageMaker returns this token. To retrieve the next set of tags, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

ResourceArn

The Amazon Resource Name (ARN) of the resource whose tags you want to retrieve.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z-]*:sagemaker:[a-z0-9-]*:[0-9]{12}:.+`

Required: Yes

Response Syntax

```
{
   "NextToken": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If response is truncated, SageMaker includes a token in the response. You can use this token in your subsequent request to fetch next set of tokens.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Tags

An array of Tag objects, each with a tag key and a value.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListTrainingJobs

Service: Amazon SageMaker Service

Lists training jobs.

> **Note**
>
> When `StatusEquals` and `MaxResults` are set at the same time, the `MaxResults` number of training jobs are first retrieved ignoring the `StatusEquals` parameter and then they are filtered by the `StatusEquals` parameter, which is returned as a response.
> For example, if `ListTrainingJobs` is invoked with the following parameters:
> `{ ... MaxResults: 100, StatusEquals: InProgress ... }`
> First, 100 trainings jobs with any status, including those other than `InProgress`, are selected (sorted according to the creation time, from the most current to the oldest). Next, those with a status of `InProgress` are returned.
> You can quickly test the API using the following AWS CLI code.
> `aws sagemaker list-training-jobs --max-results 100 --status-equals InProgress`

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string",
   "TrainingPlanArnEquals": "string",
   "WarmPoolStatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only training jobs created after the specified time (timestamp).

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only training jobs created before the specified time (timestamp).

Type: Timestamp

Required: No

LastModifiedTimeAfter

A filter that returns only training jobs modified after the specified time (timestamp).

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only training jobs modified before the specified time (timestamp).

Type: Timestamp

Required: No

MaxResults

The maximum number of training jobs to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the training job name. This filter returns only training jobs whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the result of the previous `ListTrainingJobs` request was truncated, the response includes a `NextToken`. To retrieve the next set of training jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that retrieves only training jobs with a specific status.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting`

Required: No

TrainingPlanArnEquals

The Amazon Resource Name (ARN); of the training plan to filter training jobs by. For more information about reserving GPU capacity for your SageMaker training jobs using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Required: No

WarmPoolStatusEquals

A filter that retrieves only training jobs with a specific warm pool status.

Type: String

Valid Values: `Available | Terminated | Reused | InUse`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "TrainingJobSummaries": [
      {
         "CreationTime": number,
         "LastModifiedTime": number,
         "SecondaryStatus": "string",
         "TrainingEndTime": number,
         "TrainingJobArn": "string",
         "TrainingJobName": "string",
         "TrainingJobStatus": "string",
         "TrainingPlanArn": "string",
```

```
         "WarmPoolStatus": {
            "ResourceRetainedBillableTimeInSeconds": number,
            "ReusedByJob": "string",
            "Status": "string"
         }
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the response is truncated, SageMaker returns this token. To retrieve the next set of training jobs, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

TrainingJobSummaries

An array of `TrainingJobSummary` objects, each listing a training job.

Type: Array of TrainingJobSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListTrainingJobsForHyperParameterTuningJob

Service: Amazon SageMaker Service

Gets a list of TrainingJobSummary objects that describe the training jobs that a hyperparameter tuning job launched.

Request Syntax

```
{
   "HyperParameterTuningJobName": "string",
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HyperParameterTuningJobName

The name of the tuning job whose training jobs you want to list.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

MaxResults

The maximum number of training jobs to return. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the result of the previous `ListTrainingJobsForHyperParameterTuningJob` request was truncated, the response includes a `NextToken`. To retrieve the next set of training jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `Name`.

If the value of this field is `FinalObjectiveMetricValue`, any training jobs that did not return an objective metric are not listed.

Type: String

Valid Values: `Name | CreationTime | Status | FinalObjectiveMetricValue`

Required: No

SortOrder

The sort order for results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that returns only training jobs with the specified status.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "TrainingJobSummaries": [
      {
         "CreationTime": number,
         "FailureReason": "string",
         "FinalHyperParameterTuningJobObjectiveMetric": {
            "MetricName": "string",
            "Type": "string",
            "Value": number
         },
         "ObjectiveStatus": "string",
         "TrainingEndTime": number,
         "TrainingJobArn": "string",
         "TrainingJobDefinitionName": "string",
         "TrainingJobName": "string",
         "TrainingJobStatus": "string",
         "TrainingStartTime": number,
         "TunedHyperParameters": {
            "string" : "string"
         },
         "TuningJobName": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the result of this `ListTrainingJobsForHyperParameterTuningJob` request was truncated, the response includes a `NextToken`. To retrieve the next set of training jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

TrainingJobSummaries

A list of TrainingJobSummary objects that describe the training jobs that the `ListTrainingJobsForHyperParameterTuningJob` request returned.

Type: Array of HyperParameterTrainingJobSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListTrainingPlans

Service: Amazon SageMaker Service

Retrieves a list of training plans for the current account.

Request Syntax

```
{
   "Filters": [
      {
         "Name": "string",
         "Value": "string"
      }
   ],
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StartTimeAfter": number,
   "StartTimeBefore": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Filters

Additional filters to apply to the list of training plans.

Type: Array of TrainingPlanFilter objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

MaxResults

The maximum number of results to return in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

A token to continue pagination if more results are available.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The training plan field to sort the results by (e.g., StartTime, Status).

Type: String

Valid Values: `TrainingPlanName | StartTime | Status`

Required: No

SortOrder

The order to sort the results (Ascending or Descending).

Type: String

Valid Values: `Ascending | Descending`

Required: No

StartTimeAfter

Filter to list only training plans with an actual start time after this date.

Type: Timestamp

Required: No

StartTimeBefore

Filter to list only training plans with an actual start time before this date.

Type: Timestamp

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "TrainingPlanSummaries": [
      {
         "AvailableInstanceCount": number,
         "CurrencyCode": "string",
         "DurationHours": number,
         "DurationMinutes": number,
         "EndTime": number,
         "InUseInstanceCount": number,
         "ReservedCapacitySummaries": [
            {
               "AvailabilityZone": "string",
               "AvailabilityZoneId": "string",
               "DurationHours": number,
               "DurationMinutes": number,
               "EndTime": number,
               "InstanceType": "string",
               "ReservedCapacityArn": "string",
               "ReservedCapacityType": "string",
               "StartTime": number,
               "Status": "string",
               "TotalInstanceCount": number,
               "UltraServerCount": number,
               "UltraServerType": "string"
            }
         ],
         "StartTime": number,
         "Status": "string",
         "StatusMessage": "string",
         "TargetResources": [ "string" ],
         "TotalInstanceCount": number,
         "TotalUltraServerCount": number,
         "TrainingPlanArn": "string",
         "TrainingPlanName": "string",
         "UpfrontFee": "string"
      }
```

```
    ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

A token to continue pagination if more results are available.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

TrainingPlanSummaries

A list of summary information for the training plans.

Type: Array of TrainingPlanSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListTransformJobs

Service: Amazon SageMaker Service

Lists transform jobs.

Request Syntax

```
{
   "CreationTimeAfter": number,
   "CreationTimeBefore": number,
   "LastModifiedTimeAfter": number,
   "LastModifiedTimeBefore": number,
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreationTimeAfter

A filter that returns only transform jobs created after the specified time.

Type: Timestamp

Required: No

CreationTimeBefore

A filter that returns only transform jobs created before the specified time.

Type: Timestamp

Required: No

LastModifiedTimeAfter

A filter that returns only transform jobs modified after the specified time.

Type: Timestamp

Required: No

LastModifiedTimeBefore

A filter that returns only transform jobs modified before the specified time.

Type: Timestamp

Required: No

MaxResults

The maximum number of transform jobs to return in the response. The default value is `10`.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the transform job name. This filter returns only transform jobs whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9\-]+`

Required: No

NextToken

If the result of the previous `ListTransformJobs` request was truncated, the response includes a `NextToken`. To retrieve the next set of transform jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime | Status`

Required: No

SortOrder

The sort order for results. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

StatusEquals

A filter that retrieves only transform jobs with a specific status.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "TransformJobSummaries": [
      {
         "CreationTime": number,
         "FailureReason": "string",
         "LastModifiedTime": number,
         "TransformEndTime": number,
         "TransformJobArn": "string",
         "TransformJobName": "string",
```

```
            "TransformJobStatus": "string"
        }
    ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the response is truncated, Amazon SageMaker returns this token. To retrieve the next set of transform jobs, use it in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

TransformJobSummaries

An array of `TransformJobSummary` objects.

Type: Array of TransformJobSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListTrialComponents

Service: Amazon SageMaker Service

Lists the trial components in your account. You can sort the list by trial component name or creation time. You can filter the list to show only components that were created in a specific time range. You can also filter on one of the following:

- `ExperimentName`
- `SourceArn`
- `TrialName`

Request Syntax

```
{
   "CreatedAfter": number,
   "CreatedBefore": number,
   "ExperimentName": "string",
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "SourceArn": "string",
   "TrialName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreatedAfter

A filter that returns only components created after the specified time.

Type: Timestamp

Required: No

CreatedBefore

A filter that returns only components created before the specified time.

Type: Timestamp

Required: No

ExperimentName

A filter that returns only components that are part of the specified experiment. If you specify `ExperimentName`, you can't filter by `SourceArn` or `TrialName`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

MaxResults

The maximum number of components to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous call to `ListTrialComponents` didn't return the full set of components, the call returns a token for getting the next set of components.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The property used to sort results. The default value is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order. The default value is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

SourceArn

A filter that returns only components that have the specified source Amazon Resource Name (ARN). If you specify `SourceArn`, you can't filter by `ExperimentName` or `TrialName`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

TrialName

A filter that returns only components that are part of the specified trial. If you specify `TrialName`, you can't filter by `ExperimentName` or `SourceArn`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "TrialComponentSummaries": [
      {
```

```
         "CreatedBy": {
            "DomainId": "string",
            "IamIdentity": {
               "Arn": "string",
               "PrincipalId": "string",
               "SourceIdentity": "string"
            },
            "UserProfileArn": "string",
            "UserProfileName": "string"
         },
         "CreationTime": number,
         "DisplayName": "string",
         "EndTime": number,
         "LastModifiedBy": {
            "DomainId": "string",
            "IamIdentity": {
               "Arn": "string",
               "PrincipalId": "string",
               "SourceIdentity": "string"
            },
            "UserProfileArn": "string",
            "UserProfileName": "string"
         },
         "LastModifiedTime": number,
         "StartTime": number,
         "Status": {
            "Message": "string",
            "PrimaryStatus": "string"
         },
         "TrialComponentArn": "string",
         "TrialComponentName": "string",
         "TrialComponentSource": {
            "SourceArn": "string",
            "SourceType": "string"
         }
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

A token for getting the next set of components, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

TrialComponentSummaries

A list of the summaries of your trial components.

Type: Array of TrialComponentSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

ListTrials

Service: Amazon SageMaker Service

Lists the trials in your account. Specify an experiment name to limit the list to the trials that are part of that experiment. Specify a trial component name to limit the list to the trials that associated with that trial component. The list can be filtered to show only trials that were created in a specific time range. The list can be sorted by trial name or creation time.

Request Syntax

```
{
   "CreatedAfter": number,
   "CreatedBefore": number,
   "ExperimentName": "string",
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "TrialComponentName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CreatedAfter

A filter that returns only trials created after the specified time.

Type: Timestamp

Required: No

CreatedBefore

A filter that returns only trials created before the specified time.

Type: Timestamp

Required: No

ExperimentName

A filter that returns only trials that are part of the specified experiment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

MaxResults

The maximum number of trials to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous call to `ListTrials` didn't return the full set of trials, the call returns a token for getting the next set of trials.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The property used to sort results. The default value is `CreationTime`.

Type: String

Valid Values: `Name | CreationTime`

Required: No

SortOrder

The sort order. The default value is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

TrialComponentName

A filter that returns only trials that are associated with the specified trial component.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "TrialSummaries": [
      {
         "CreationTime": number,
         "DisplayName": "string",
         "LastModifiedTime": number,
         "TrialArn": "string",
         "TrialName": "string",
         "TrialSource": {
            "SourceArn": "string",
            "SourceType": "string"
         }
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

A token for getting the next set of trials, if there are any.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

TrialSummaries

A list of the summaries of your trials.

Type: Array of TrialSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

ListUltraServersByReservedCapacity

Service: Amazon SageMaker Service

Lists all UltraServers that are part of a specified reserved capacity.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string",
   "ReservedCapacityArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

The maximum number of UltraServers to return in the response. The default value is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

ReservedCapacityArn

The ARN of the reserved capacity to list UltraServers for.

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:reserved-capacity/.*`

Required: Yes

Response Syntax

```
{
   "NextToken": "string",
   "UltraServers": [
      {
         "AvailabilityZone": "string",
         "AvailableInstanceCount": number,
         "AvailableSpareInstanceCount": number,
         "ConfiguredSpareInstanceCount": number,
         "HealthStatus": "string",
         "InstanceType": "string",
         "InUseInstanceCount": number,
         "TotalInstanceCount": number,
         "UltraServerId": "string",
         "UltraServerType": "string",
         "UnhealthyInstanceCount": number
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the response is truncated, SageMaker returns this token. Use it in the next request to retrieve the next set of UltraServers.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

UltraServers

A list of UltraServers that are part of the specified reserved capacity.

Type: Array of UltraServer objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListUserProfiles

Service: Amazon SageMaker Service

Lists user profiles.

Request Syntax

```
{
   "DomainIdEquals": "string",
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "UserProfileNameContains": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainIdEquals

A parameter by which to filter the results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: No

MaxResults

This parameter defines the maximum number of results that can be return in a single response. The `MaxResults` parameter is an upper bound, not a target. If there are more results available than the value specified, a `NextToken` is provided in the response. The `NextToken` indicates that the user should get the next set of results by providing this token as a part of a subsequent call. The default value for `MaxResults` is 10.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The parameter by which to sort the results. The default is CreationTime.

Type: String

Valid Values: `CreationTime | LastModifiedTime`

Required: No

SortOrder

The sort order for the results. The default is Ascending.

Type: String

Valid Values: `Ascending | Descending`

Required: No

UserProfileNameContains

A parameter by which to filter the results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "UserProfiles": [
      {
         "CreationTime": number,
         "DomainId": "string",
         "LastModifiedTime": number,
         "Status": "string",
         "UserProfileName": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the previous response was truncated, you will receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

UserProfiles

The list of user profiles.

Type: Array of UserProfileDetails objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListWorkforces

Service: Amazon SageMaker Service

Use this operation to list all private and vendor workforces in an AWS Region. Note that you can only have one private workforce per AWS Region.

Request Syntax

```
{
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

The maximum number of workforces returned in the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A filter you can use to search for workforces using part of the workforce name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9\-]){0,62}`

Required: No

NextToken

A token to resume pagination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

Sort workforces using the workforce name or creation date.

Type: String

Valid Values: `Name | CreateDate`

Required: No

SortOrder

Sort workforces in ascending or descending order.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "Workforces": [
      {
         "CognitoConfig": {
            "ClientId": "string",
            "UserPool": "string"
         },
         "CreateDate": number,
         "FailureReason": "string",
         "IpAddressType": "string",
```

```
         "LastUpdatedDate": number,
         "OidcConfig": {
            "AuthenticationRequestExtraParams": {
               "string" : "string"
            },
            "AuthorizationEndpoint": "string",
            "ClientId": "string",
            "Issuer": "string",
            "JwksUri": "string",
            "LogoutEndpoint": "string",
            "Scope": "string",
            "TokenEndpoint": "string",
            "UserInfoEndpoint": "string"
         },
         "SourceIpConfig": {
            "Cidrs": [ "string" ]
         },
         "Status": "string",
         "SubDomain": "string",
         "WorkforceArn": "string",
         "WorkforceName": "string",
         "WorkforceVpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ],
            "VpcEndpointId": "string",
            "VpcId": "string"
         }
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

A token to resume pagination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Workforces

A list containing information about your workforce.

Type: Array of Workforce objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListWorkteams

Service: Amazon SageMaker Service

Gets a list of private work teams that you have defined in a region. The list may be empty if no work team satisfies the filter specified in the `NameContains` parameter.

Request Syntax

```
{
   "MaxResults": number,
   "NameContains": "string",
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MaxResults

The maximum number of work teams to return in each page of the response.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NameContains

A string in the work team's name. This filter returns only work teams whose name contains the specified string.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

NextToken

If the result of the previous `ListWorkteams` request was truncated, the response includes a `NextToken`. To retrieve the next set of labeling jobs, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

SortBy

The field to sort results by. The default is `CreationTime`.

Type: String

Valid Values: `Name | CreateDate`

Required: No

SortOrder

The sort order for results. The default is `Ascending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "Workteams": [
      {
         "CreateDate": number,
         "Description": "string",
         "LastUpdatedDate": number,
         "MemberDefinitions": [
```

```
            {
               "CognitoMemberDefinition": {
                  "ClientId": "string",
                  "UserGroup": "string",
                  "UserPool": "string"
               },
               "OidcMemberDefinition": {
                  "Groups": [ "string" ]
               }
            }
         ],
         "NotificationConfiguration": {
            "NotificationTopicArn": "string"
         },
         "ProductListingIds": [ "string" ],
         "SubDomain": "string",
         "WorkerAccessConfiguration": {
            "S3Presign": {
               "IamPolicyConstraints": {
                  "SourceIp": "string",
                  "VpcSourceIp": "string"
               }
            }
         },
         "WorkforceArn": "string",
         "WorkteamArn": "string",
         "WorkteamName": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the response is truncated, Amazon SageMaker returns this token. To retrieve the next set of work teams, use it in the subsequent request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Workteams

An array of `Workteam` objects, each describing a work team.

Type: Array of Workteam objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

PutModelPackageGroupPolicy

Service: Amazon SageMaker Service

Adds a resouce policy to control access to a model group. For information about resoure policies, see Identity-based policies and resource-based policies in the *AWS Identity and Access Management User Guide.*.

Request Syntax

```
{
   "ModelPackageGroupName": "string",
   "ResourcePolicy": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ModelPackageGroupName

The name of the model group to add a resource policy to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ResourcePolicy

The resource policy for the model group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20480.

Pattern: `.*`

Required: Yes

Response Syntax

```
{
   "ModelPackageGroupArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelPackageGroupArn

The Amazon Resource Name (ARN) of the model package group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package-group/[\S]{1,2048}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

QueryLineage

Service: Amazon SageMaker Service

Use this action to inspect your lineage and discover relationships between entities. For more information, see Querying Lineage Entities in the *Amazon SageMaker Developer Guide*.

Request Syntax

```
{
   "Direction": "string",
   "Filters": {
      "CreatedAfter": number,
      "CreatedBefore": number,
      "LineageTypes": [ "string" ],
      "ModifiedAfter": number,
      "ModifiedBefore": number,
      "Properties": {
         "string" : "string"
      },
      "Types": [ "string" ]
   },
   "IncludeEdges": boolean,
   "MaxDepth": number,
   "MaxResults": number,
   "NextToken": "string",
   "StartArns": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Direction

Associations between lineage entities have a direction. This parameter determines the direction from the StartArn(s) that the query traverses.

Type: String

Valid Values: `Both | Ascendants | Descendants`

Required: No

Filters

A set of filtering parameters that allow you to specify which entities should be returned.

- Properties - Key-value pairs to match on the lineage entities' properties.
- LineageTypes - A set of lineage entity types to match on. For example: `TrialComponent`, `Artifact`, or `Context`.
- CreatedBefore - Filter entities created before this date.
- ModifiedBefore - Filter entities modified before this date.
- ModifiedAfter - Filter entities modified after this date.

Type: QueryFilters object

Required: No

IncludeEdges

Setting this value to `True` retrieves not only the entities of interest but also the Associations and lineage entities on the path. Set to `False` to only return lineage entities that match your query.

Type: Boolean

Required: No

MaxDepth

The maximum depth in lineage relationships from the `StartArns` that are traversed. Depth is a measure of the number of `Associations` from the `StartArn` entity to the matched results.

Type: Integer

Valid Range: Maximum value of 10.

Required: No

MaxResults

Limits the number of vertices in the results. Use the `NextToken` in a response to to retrieve the next page of results.

Type: Integer

Valid Range: Maximum value of 50.

Required: No

NextToken

Limits the number of vertices in the request. Use the `NextToken` in a response to to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Required: No

StartArns

A list of resource Amazon Resource Name (ARN) that represent the starting point for your lineage query.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: No

Response Syntax

```
{
   "Edges": [
      {
         "AssociationType": "string",
         "DestinationArn": "string",
         "SourceArn": "string"
      }
   ],
   "NextToken": "string",
   "Vertices": [
```

```
      {
         "Arn": "string",
         "LineageType": "string",
         "Type": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Edges

A list of edges that connect vertices in the response.

Type: Array of Edge objects

NextToken

Limits the number of vertices in the response. Use the `NextToken` in a response to to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Vertices

A list of vertices connected to the start entity(ies) in the lineage graph.

Type: Array of Vertex objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

RegisterDevices

Service: Amazon SageMaker Service

Register devices.

Request Syntax

```
{
   "DeviceFleetName": "string",
   "Devices": [
      {
         "Description": "string",
         "DeviceName": "string",
         "IotThingName": "string"
      }
   ],
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeviceFleetName

The name of the fleet.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Devices

A list of devices to register with SageMaker Edge Manager.

Type: Array of Device objects

Required: Yes

Tags

The tags associated with devices.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

RenderUiTemplate

Service: Amazon SageMaker Service

Renders the UI template so that you can preview the worker's experience.

Request Syntax

```
{
   "HumanTaskUiArn": "string",
   "RoleArn": "string",
   "Task": {
      "Input": "string"
   },
   "UiTemplate": {
      "Content": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HumanTaskUiArn

The `HumanTaskUiArn` of the worker UI that you want to render. Do not provide a `HumanTaskUiArn` if you use the `UiTemplate` parameter.

See a list of available Human Ui Amazon Resource Names (ARNs) in UiConfig.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]+:[0-9]{12}:human-task-ui/.*`

Required: No

RoleArn

The Amazon Resource Name (ARN) that has access to the S3 objects that are used by the template.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Task

A `RenderableTask` object containing a representative task to render.

Type: RenderableTask object

Required: Yes

UiTemplate

A `Template` object containing the worker UI template to render.

Type: UiTemplate object

Required: No

Response Syntax

```
{
   "Errors": [
      {
         "Code": "string",
         "Message": "string"
      }
   ],
   "RenderedContent": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Errors

A list of one or more `RenderingError` objects if any were encountered while rendering the template. If there were no errors, the list is empty.

Type: Array of RenderingError objects

RenderedContent

A Liquid template that renders the HTML for the worker UI.

Type: String

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

RetryPipelineExecution

Service: Amazon SageMaker Service

Retry the execution of the pipeline.

Request Syntax

```
{
   "ClientRequestToken": "string",
   "ParallelismConfiguration": {
      "MaxParallelExecutionSteps": number
   },
   "PipelineExecutionArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClientRequestToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the operation. An idempotent operation completes no more than once.

Type: String

Length Constraints: Minimum length of 32. Maximum length of 128.

Required: Yes

ParallelismConfiguration

This configuration, if specified, overrides the parallelism configuration of the parent pipeline.

Type: ParallelismConfiguration object

Required: No

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: Yes

Response Syntax

```
{
   "PipelineExecutionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Search

Service: Amazon SageMaker Service

Finds SageMaker resources that match a search query. Matching resources are returned as a list of `SearchRecord` objects in the response. You can sort the search results by any resource property in a ascending or descending order.

You can query against the following value types: numeric, text, Boolean, and timestamp.

> **Note**
>
> The Search API may provide access to otherwise restricted data. See Amazon SageMaker API Permissions: Actions, Permissions, and Resources Reference for more information.

Request Syntax

```
{
   "CrossAccountFilterOption": "string",
   "MaxResults": number,
   "NextToken": "string",
   "Resource": "string",
   "SearchExpression": {
      "Filters": [
         {
            "Name": "string",
            "Operator": "string",
            "Value": "string"
         }
      ],
      "NestedFilters": [
         {
            "Filters": [
               {
                  "Name": "string",
                  "Operator": "string",
                  "Value": "string"
               }
            ],
            "NestedPropertyName": "string"
         }
      ],
```

```
      "Operator": "string",
      "SubExpressions": [
         "SearchExpression"
      ]
   },
   "SortBy": "string",
   "SortOrder": "string",
   "VisibilityConditions": [
      {
         "Key": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CrossAccountFilterOption

A cross account filter option. When the value is "CrossAccount" the search results will only include resources made discoverable to you from other accounts. When the value is "SameAccount" or null the search results will only include resources from your account. Default is null. For more information on searching for resources made discoverable to your account, see Search discoverable resources in the SageMaker Developer Guide. The maximum number of ResourceCatalogs viewable is 1000.

Type: String

Valid Values: SameAccount | CrossAccount

Required: No

MaxResults

The maximum number of results to return.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

NextToken

If more than `MaxResults` resources match the specified `SearchExpression`, the response includes a `NextToken`. The `NextToken` can be passed to the next `SearchRequest` to continue retrieving results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Required: No

Resource

The name of the SageMaker resource to search for.

Type: String

Valid Values: `TrainingJob | Experiment | ExperimentTrial | ExperimentTrialComponent | Endpoint | Model | ModelPackage | ModelPackageGroup | Pipeline | PipelineExecution | FeatureGroup | FeatureMetadata | Image | ImageVersion | Project | HyperParameterTuningJob | ModelCard | PipelineVersion | Job`

Required: Yes

SearchExpression

A Boolean conditional statement. Resources must satisfy this condition to be included in search results. You must provide at least one subexpression, filter, or nested filter. The maximum number of recursive `SubExpressions`, `NestedFilters`, and `Filters` that can be included in a `SearchExpression` object is 50.

Type: SearchExpression object

Required: No

SortBy

The name of the resource property used to sort the `SearchResults`. The default is `LastModifiedTime`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: No

SortOrder

How `SearchResults` are ordered. Valid values are `Ascending` or `Descending`. The default is `Descending`.

Type: String

Valid Values: `Ascending | Descending`

Required: No

VisibilityConditions

Limits the results of your search request to the resources that you can access.

Type: Array of VisibilityConditions objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "Results": [
      {
         "Endpoint": {
            "CreationTime": number,
            "DataCaptureConfig": {
               "CaptureStatus": "string",
               "CurrentSamplingPercentage": number,
               "DestinationS3Uri": "string",
               "EnableCapture": boolean,
               "KmsKeyId": "string"
            },
            "EndpointArn": "string",
```

```
            "EndpointConfigName": "string",
            "EndpointName": "string",
            "EndpointStatus": "string",
            "FailureReason": "string",
            "LastModifiedTime": number,
            "MonitoringSchedules": [
               {
                  "CreationTime": number,
                  "EndpointName": "string",
                  "FailureReason": "string",
                  "LastModifiedTime": number,
                  "LastMonitoringExecutionSummary": {
                     "CreationTime": number,
                     "EndpointName": "string",
                     "FailureReason": "string",
                     "LastModifiedTime": number,
                     "MonitoringExecutionStatus": "string",
                     "MonitoringJobDefinitionName": "string",
                     "MonitoringScheduleName": "string",
                     "MonitoringType": "string",
                     "ProcessingJobArn": "string",
                     "ScheduledTime": number
                  },
                  "MonitoringScheduleArn": "string",
                  "MonitoringScheduleConfig": {
                     "MonitoringJobDefinition": {
                        "BaselineConfig": {
                           "BaseliningJobName": "string",
                           "ConstraintsResource": {
                              "S3Uri": "string"
                           },
                           "StatisticsResource": {
                              "S3Uri": "string"
                           }
                        },
                        "Environment": {
                           "string" : "string"
                        },
                        "MonitoringAppSpecification": {
                           "ContainerArguments": [ "string" ],
                           "ContainerEntrypoint": [ "string" ],
                           "ImageUri": "string",
                           "PostAnalyticsProcessorSourceUri": "string",
                           "RecordPreprocessorSourceUri": "string"
```

```
        },
        "MonitoringInputs": [
           {
              "BatchTransformInput": {
                 "DataCapturedDestinationS3Uri": "string",
                 "DatasetFormat": {
                    "Csv": {
                       "Header": boolean
                    },
                    "Json": {
                       "Line": boolean
                    },
                    "Parquet": {
                    }
                 },
                 "EndTimeOffset": "string",
                 "ExcludeFeaturesAttribute": "string",
                 "FeaturesAttribute": "string",
                 "InferenceAttribute": "string",
                 "LocalPath": "string",
                 "ProbabilityAttribute": "string",
                 "ProbabilityThresholdAttribute": number,
                 "S3DataDistributionType": "string",
                 "S3InputMode": "string",
                 "StartTimeOffset": "string"
              },
              "EndpointInput": {
                 "EndpointName": "string",
                 "EndTimeOffset": "string",
                 "ExcludeFeaturesAttribute": "string",
                 "FeaturesAttribute": "string",
                 "InferenceAttribute": "string",
                 "LocalPath": "string",
                 "ProbabilityAttribute": "string",
                 "ProbabilityThresholdAttribute": number,
                 "S3DataDistributionType": "string",
                 "S3InputMode": "string",
                 "StartTimeOffset": "string"
              }
           }
        ],
        "MonitoringOutputConfig": {
           "KmsKeyId": "string",
           "MonitoringOutputs": [
```

```
                        {
                           "S3Output": {
                              "LocalPath": "string",
                              "S3UploadMode": "string",
                              "S3Uri": "string"
                           }
                        }
                     ]
                  },
                  "MonitoringResources": {
                     "ClusterConfig": {
                        "InstanceCount": number,
                        "InstanceType": "string",
                        "VolumeKmsKeyId": "string",
                        "VolumeSizeInGB": number
                     }
                  },
                  "NetworkConfig": {
                     "EnableInterContainerTrafficEncryption": boolean,
                     "EnableNetworkIsolation": boolean,
                     "VpcConfig": {
                        "SecurityGroupIds": [ "string" ],
                        "Subnets": [ "string" ]
                     }
                  },
                  "RoleArn": "string",
                  "StoppingCondition": {
                     "MaxRuntimeInSeconds": number
                  }
               },
               "MonitoringJobDefinitionName": "string",
               "MonitoringType": "string",
               "ScheduleConfig": {
                  "DataAnalysisEndTime": "string",
                  "DataAnalysisStartTime": "string",
                  "ScheduleExpression": "string"
               }
            },
            "MonitoringScheduleName": "string",
            "MonitoringScheduleStatus": "string",
            "MonitoringType": "string",
            "Tags": [
               {
                  "Key": "string",
```

```
                        "Value": "string"
                    }
                ]
            }
        ],
        "ProductionVariants": [
            {
                "CapacityReservationConfig": {
                    "AvailableInstanceCount": number,
                    "CapacityReservationPreference": "string",
                    "Ec2CapacityReservations": [
                        {
                            "AvailableInstanceCount": number,
                            "Ec2CapacityReservationId": "string",
                            "TotalInstanceCount": number,
                            "UsedByCurrentEndpoint": number
                        }
                    ],
                    "MlReservationArn": "string",
                    "TotalInstanceCount": number,
                    "UsedByCurrentEndpoint": number
                },
                "CurrentInstanceCount": number,
                "CurrentServerlessConfig": {
                    "MaxConcurrency": number,
                    "MemorySizeInMB": number,
                    "ProvisionedConcurrency": number
                },
                "CurrentWeight": number,
                "DeployedImages": [
                    {
                        "ResolutionTime": number,
                        "ResolvedImage": "string",
                        "SpecifiedImage": "string"
                    }
                ],
                "DesiredInstanceCount": number,
                "DesiredServerlessConfig": {
                    "MaxConcurrency": number,
                    "MemorySizeInMB": number,
                    "ProvisionedConcurrency": number
                },
                "DesiredWeight": number,
                "InstancePools": [
```

```
                {
                   "CurrentInstanceCount": number,
                   "InstanceType": "string"
                }
             ],
             "ManagedInstanceScaling": {
                "MaxInstanceCount": number,
                "MinInstanceCount": number,
                "ScaleInPolicy": {
                   "CooldownInMinutes": number,
                   "MaximumStepSize": number,
                   "Strategy": "string"
                },
                "Status": "string"
             },
             "RoutingConfig": {
                "RoutingStrategy": "string"
             },
             "VariantName": "string",
             "VariantStatus": [
                {
                   "StartTime": number,
                   "Status": "string",
                   "StatusMessage": "string"
                }
             ]
          }
       ],
       "ShadowProductionVariants": [
          {
             "CapacityReservationConfig": {
                "AvailableInstanceCount": number,
                "CapacityReservationPreference": "string",
                "Ec2CapacityReservations": [
                   {
                      "AvailableInstanceCount": number,
                      "Ec2CapacityReservationId": "string",
                      "TotalInstanceCount": number,
                      "UsedByCurrentEndpoint": number
                   }
                ],
                "MlReservationArn": "string",
                "TotalInstanceCount": number,
                "UsedByCurrentEndpoint": number
```

```
      },
      "CurrentInstanceCount": number,
      "CurrentServerlessConfig": {
         "MaxConcurrency": number,
         "MemorySizeInMB": number,
         "ProvisionedConcurrency": number
      },
      "CurrentWeight": number,
      "DeployedImages": [
         {
            "ResolutionTime": number,
            "ResolvedImage": "string",
            "SpecifiedImage": "string"
         }
      ],
      "DesiredInstanceCount": number,
      "DesiredServerlessConfig": {
         "MaxConcurrency": number,
         "MemorySizeInMB": number,
         "ProvisionedConcurrency": number
      },
      "DesiredWeight": number,
      "InstancePools": [
         {
            "CurrentInstanceCount": number,
            "InstanceType": "string"
         }
      ],
      "ManagedInstanceScaling": {
         "MaxInstanceCount": number,
         "MinInstanceCount": number,
         "ScaleInPolicy": {
            "CooldownInMinutes": number,
            "MaximumStepSize": number,
            "Strategy": "string"
         },
         "Status": "string"
      },
      "RoutingConfig": {
         "RoutingStrategy": "string"
      },
      "VariantName": "string",
      "VariantStatus": [
         {
```

```
                        "StartTime": number,
                        "Status": "string",
                        "StatusMessage": "string"
                     }
                  ]
               }
            ],
            "Tags": [
               {
                  "Key": "string",
                  "Value": "string"
               }
            ]
         },
         "Experiment": {
            "CreatedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "CreationTime": number,
            "Description": "string",
            "DisplayName": "string",
            "ExperimentArn": "string",
            "ExperimentName": "string",
            "LastModifiedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "LastModifiedTime": number,
            "Source": {
               "SourceArn": "string",
               "SourceType": "string"
```

```
            },
            "Tags": [
               {
                  "Key": "string",
                  "Value": "string"
               }
            ]
         },
         "FeatureGroup": {
            "CreationTime": number,
            "Description": "string",
            "EventTimeFeatureName": "string",
            "FailureReason": "string",
            "FeatureDefinitions": [
               {
                  "CollectionConfig": { ... },
                  "CollectionType": "string",
                  "FeatureName": "string",
                  "FeatureType": "string"
               }
            ],
            "FeatureGroupArn": "string",
            "FeatureGroupName": "string",
            "FeatureGroupStatus": "string",
            "LastModifiedTime": number,
            "LastUpdateStatus": {
               "FailureReason": "string",
               "Status": "string"
            },
            "OfflineStoreConfig": {
               "DataCatalogConfig": {
                  "Catalog": "string",
                  "Database": "string",
                  "TableName": "string"
               },
               "DisableGlueTableCreation": boolean,
               "S3StorageConfig": {
                  "KmsKeyId": "string",
                  "ResolvedOutputS3Uri": "string",
                  "S3Uri": "string"
               },
               "TableFormat": "string"
            },
            "OfflineStoreStatus": {
```

```
            "BlockedReason": "string",
            "Status": "string"
         },
         "OnlineStoreConfig": {
            "EnableOnlineStore": boolean,
            "SecurityConfig": {
               "KmsKeyId": "string"
            },
            "StorageType": "string",
            "TtlDuration": {
               "Unit": "string",
               "Value": number
            }
         },
         "RecordIdentifierFeatureName": "string",
         "RoleArn": "string",
         "Tags": [
            {
               "Key": "string",
               "Value": "string"
            }
         ]
      },
      "FeatureMetadata": {
         "CreationTime": number,
         "Description": "string",
         "FeatureGroupArn": "string",
         "FeatureGroupName": "string",
         "FeatureName": "string",
         "FeatureType": "string",
         "LastModifiedTime": number,
         "Parameters": [
            {
               "Key": "string",
               "Value": "string"
            }
         ]
      },
      "HyperParameterTuningJob": {
         "BestTrainingJob": {
            "CreationTime": number,
            "FailureReason": "string",
            "FinalHyperParameterTuningJobObjectiveMetric": {
               "MetricName": "string",
```

```
            "Type": "string",
            "Value": number
         },
         "ObjectiveStatus": "string",
         "TrainingEndTime": number,
         "TrainingJobArn": "string",
         "TrainingJobDefinitionName": "string",
         "TrainingJobName": "string",
         "TrainingJobStatus": "string",
         "TrainingStartTime": number,
         "TunedHyperParameters": {
            "string" : "string"
         },
         "TuningJobName": "string"
      },
      "ConsumedResources": {
         "RuntimeInSeconds": number
      },
      "CreationTime": number,
      "FailureReason": "string",
      "HyperParameterTuningEndTime": number,
      "HyperParameterTuningJobArn": "string",
      "HyperParameterTuningJobConfig": {
         "HyperParameterTuningJobObjective": {
            "MetricName": "string",
            "Type": "string"
         },
         "ParameterRanges": {
            "AutoParameters": [
               {
                  "Name": "string",
                  "ValueHint": "string"
               }
            ],
            "CategoricalParameterRanges": [
               {
                  "Name": "string",
                  "Values": [ "string" ]
               }
            ],
            "ContinuousParameterRanges": [
               {
                  "MaxValue": "string",
                  "MinValue": "string",
```

```
                        "Name": "string",
                        "ScalingType": "string"
                     }
                  ],
                  "IntegerParameterRanges": [
                     {
                        "MaxValue": "string",
                        "MinValue": "string",
                        "Name": "string",
                        "ScalingType": "string"
                     }
                  ]
               },
               "RandomSeed": number,
               "ResourceLimits": {
                  "MaxNumberOfTrainingJobs": number,
                  "MaxParallelTrainingJobs": number,
                  "MaxRuntimeInSeconds": number
               },
               "Strategy": "string",
               "StrategyConfig": {
                  "HyperbandStrategyConfig": {
                     "MaxResource": number,
                     "MinResource": number
                  }
               },
               "TrainingJobEarlyStoppingType": "string",
               "TuningJobCompletionCriteria": {
                  "BestObjectiveNotImproving": {
                     "MaxNumberOfTrainingJobsNotImproving": number
                  },
                  "ConvergenceDetected": {
                     "CompleteOnConvergence": "string"
                  },
                  "TargetObjectiveMetricValue": number
               }
            },
            "HyperParameterTuningJobName": "string",
            "HyperParameterTuningJobStatus": "string",
            "LastModifiedTime": number,
            "ObjectiveStatusCounters": {
               "Failed": number,
               "Pending": number,
               "Succeeded": number
```

```
            },
            "OverallBestTrainingJob": {
               "CreationTime": number,
               "FailureReason": "string",
               "FinalHyperParameterTuningJobObjectiveMetric": {
                  "MetricName": "string",
                  "Type": "string",
                  "Value": number
               },
               "ObjectiveStatus": "string",
               "TrainingEndTime": number,
               "TrainingJobArn": "string",
               "TrainingJobDefinitionName": "string",
               "TrainingJobName": "string",
               "TrainingJobStatus": "string",
               "TrainingStartTime": number,
               "TunedHyperParameters": {
                  "string" : "string"
               },
               "TuningJobName": "string"
            },
            "Tags": [
               {
                  "Key": "string",
                  "Value": "string"
               }
            ],
            "TrainingJobDefinition": {
               "AlgorithmSpecification": {
                  "AlgorithmName": "string",
                  "MetricDefinitions": [
                     {
                        "Name": "string",
                        "Regex": "string"
                     }
                  ],
                  "TrainingImage": "string",
                  "TrainingInputMode": "string"
               },
               "CheckpointConfig": {
                  "LocalPath": "string",
                  "S3Uri": "string"
               },
               "DefinitionName": "string",
```

```
            "EnableInterContainerTrafficEncryption": boolean,
            "EnableManagedSpotTraining": boolean,
            "EnableNetworkIsolation": boolean,
            "Environment": {
               "string" : "string"
            },
            "HyperParameterRanges": {
               "AutoParameters": [
                  {
                     "Name": "string",
                     "ValueHint": "string"
                  }
               ],
               "CategoricalParameterRanges": [
                  {
                     "Name": "string",
                     "Values": [ "string" ]
                  }
               ],
               "ContinuousParameterRanges": [
                  {
                     "MaxValue": "string",
                     "MinValue": "string",
                     "Name": "string",
                     "ScalingType": "string"
                  }
               ],
               "IntegerParameterRanges": [
                  {
                     "MaxValue": "string",
                     "MinValue": "string",
                     "Name": "string",
                     "ScalingType": "string"
                  }
               ]
            },
            "HyperParameterTuningResourceConfig": {
               "AllocationStrategy": "string",
               "InstanceConfigs": [
                  {
                     "InstanceCount": number,
                     "InstanceType": "string",
                     "VolumeSizeInGB": number
                  }
```

```
      ],
      "InstanceCount": number,
      "InstanceType": "string",
      "VolumeKmsKeyId": "string",
      "VolumeSizeInGB": number
   },
   "InputDataConfig": [
      {
         "ChannelName": "string",
         "CompressionType": "string",
         "ContentType": "string",
         "DataSource": {
            "DatasetSource": {
               "DatasetArn": "string"
            },
            "FileSystemDataSource": {
               "DirectoryPath": "string",
               "FileSystemAccessMode": "string",
               "FileSystemId": "string",
               "FileSystemType": "string"
            },
            "S3DataSource": {
               "AttributeNames": [ "string" ],
               "HubAccessConfig": {
                  "HubContentArn": "string"
               },
               "InstanceGroupNames": [ "string" ],
               "ModelAccessConfig": {
                  "AcceptEula": boolean
               },
               "S3DataDistributionType": "string",
               "S3DataType": "string",
               "S3Uri": "string"
            }
         },
         "InputMode": "string",
         "RecordWrapperType": "string",
         "ShuffleConfig": {
            "Seed": number
         }
      }
   ],
   "OutputDataConfig": {
      "CompressionType": "string",
```

```
      "KmsKeyId": "string",
      "S3OutputPath": "string"
   },
   "ResourceConfig": {
      "InstanceCount": number,
      "InstanceGroups": [
         {
            "InstanceCount": number,
            "InstanceGroupName": "string",
            "InstanceType": "string"
         }
      ],
      "InstancePlacementConfig": {
         "EnableMultipleJobs": boolean,
         "PlacementSpecifications": [
            {
               "InstanceCount": number,
               "UltraServerId": "string"
            }
         ]
      },
      "InstanceType": "string",
      "KeepAlivePeriodInSeconds": number,
      "TrainingPlanArn": "string",
      "VolumeKmsKeyId": "string",
      "VolumeSizeInGB": number
   },
   "RetryStrategy": {
      "MaximumRetryAttempts": number
   },
   "RoleArn": "string",
   "StaticHyperParameters": {
      "string" : "string"
   },
   "StoppingCondition": {
      "MaxPendingTimeInSeconds": number,
      "MaxRuntimeInSeconds": number,
      "MaxWaitTimeInSeconds": number
   },
   "TuningObjective": {
      "MetricName": "string",
      "Type": "string"
   },
   "VpcConfig": {
```

```
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      },
      "TrainingJobDefinitions": [
         {
            "AlgorithmSpecification": {
               "AlgorithmName": "string",
               "MetricDefinitions": [
                  {
                     "Name": "string",
                     "Regex": "string"
                  }
               ],
               "TrainingImage": "string",
               "TrainingInputMode": "string"
            },
            "CheckpointConfig": {
               "LocalPath": "string",
               "S3Uri": "string"
            },
            "DefinitionName": "string",
            "EnableInterContainerTrafficEncryption": boolean,
            "EnableManagedSpotTraining": boolean,
            "EnableNetworkIsolation": boolean,
            "Environment": {
               "string" : "string"
            },
            "HyperParameterRanges": {
               "AutoParameters": [
                  {
                     "Name": "string",
                     "ValueHint": "string"
                  }
               ],
               "CategoricalParameterRanges": [
                  {
                     "Name": "string",
                     "Values": [ "string" ]
                  }
               ],
               "ContinuousParameterRanges": [
                  {
                     "MaxValue": "string",
```

```
                        "MinValue": "string",
                        "Name": "string",
                        "ScalingType": "string"
                     }
                  ],
                  "IntegerParameterRanges": [
                     {
                        "MaxValue": "string",
                        "MinValue": "string",
                        "Name": "string",
                        "ScalingType": "string"
                     }
                  ]
               },
               "HyperParameterTuningResourceConfig": {
                  "AllocationStrategy": "string",
                  "InstanceConfigs": [
                     {
                        "InstanceCount": number,
                        "InstanceType": "string",
                        "VolumeSizeInGB": number
                     }
                  ],
                  "InstanceCount": number,
                  "InstanceType": "string",
                  "VolumeKmsKeyId": "string",
                  "VolumeSizeInGB": number
               },
               "InputDataConfig": [
                  {
                     "ChannelName": "string",
                     "CompressionType": "string",
                     "ContentType": "string",
                     "DataSource": {
                        "DatasetSource": {
                           "DatasetArn": "string"
                        },
                        "FileSystemDataSource": {
                           "DirectoryPath": "string",
                           "FileSystemAccessMode": "string",
                           "FileSystemId": "string",
                           "FileSystemType": "string"
                        },
                        "S3DataSource": {
```

```
                        "AttributeNames": [ "string" ],
                        "HubAccessConfig": {
                           "HubContentArn": "string"
                        },
                        "InstanceGroupNames": [ "string" ],
                        "ModelAccessConfig": {
                           "AcceptEula": boolean
                        },
                        "S3DataDistributionType": "string",
                        "S3DataType": "string",
                        "S3Uri": "string"
                     }
                  },
                  "InputMode": "string",
                  "RecordWrapperType": "string",
                  "ShuffleConfig": {
                     "Seed": number
                  }
               }
            ],
            "OutputDataConfig": {
               "CompressionType": "string",
               "KmsKeyId": "string",
               "S3OutputPath": "string"
            },
            "ResourceConfig": {
               "InstanceCount": number,
               "InstanceGroups": [
                  {
                     "InstanceCount": number,
                     "InstanceGroupName": "string",
                     "InstanceType": "string"
                  }
               ],
               "InstancePlacementConfig": {
                  "EnableMultipleJobs": boolean,
                  "PlacementSpecifications": [
                     {
                        "InstanceCount": number,
                        "UltraServerId": "string"
                     }
                  ]
               },
               "InstanceType": "string",
```

```
                  "KeepAlivePeriodInSeconds": number,
                  "TrainingPlanArn": "string",
                  "VolumeKmsKeyId": "string",
                  "VolumeSizeInGB": number
               },
               "RetryStrategy": {
                  "MaximumRetryAttempts": number
               },
               "RoleArn": "string",
               "StaticHyperParameters": {
                  "string" : "string"
               },
               "StoppingCondition": {
                  "MaxPendingTimeInSeconds": number,
                  "MaxRuntimeInSeconds": number,
                  "MaxWaitTimeInSeconds": number
               },
               "TuningObjective": {
                  "MetricName": "string",
                  "Type": "string"
               },
               "VpcConfig": {
                  "SecurityGroupIds": [ "string" ],
                  "Subnets": [ "string" ]
               }
            }
         ],
         "TrainingJobStatusCounters": {
            "Completed": number,
            "InProgress": number,
            "NonRetryableError": number,
            "RetryableError": number,
            "Stopped": number
         },
         "TuningJobCompletionDetails": {
            "ConvergenceDetectedTime": number,
            "NumberOfTrainingJobsObjectiveNotImproving": number
         },
         "WarmStartConfig": {
            "ParentHyperParameterTuningJobs": [
               {
                  "HyperParameterTuningJobName": "string"
               }
            ],
```

```
            "WarmStartType": "string"
         }
      },
      "Job": {
         "CreationTime": number,
         "EndTime": number,
         "FailureReason": "string",
         "JobArn": "string",
         "JobCategory": "string",
         "JobConfigDocument": "string",
         "JobConfigSchemaVersion": "string",
         "JobName": "string",
         "JobStatus": "string",
         "LastModifiedTime": number,
         "RoleArn": "string",
         "SecondaryStatus": "string",
         "SecondaryStatusTransitions": [
            {
               "EndTime": number,
               "StartTime": number,
               "Status": "string",
               "StatusMessage": "string"
            }
         ],
         "Tags": [
            {
               "Key": "string",
               "Value": "string"
            }
         ]
      },
      "Model": {
         "Endpoints": [
            {
               "CreationTime": number,
               "EndpointArn": "string",
               "EndpointName": "string",
               "EndpointStatus": "string",
               "LastModifiedTime": number
            }
         ],
         "LastBatchTransformJob": {
            "AutoMLJobArn": "string",
            "BatchStrategy": "string",
```

```
   "CreationTime": number,
   "DataCaptureConfig": {
      "DestinationS3Uri": "string",
      "GenerateInferenceId": boolean,
      "KmsKeyId": "string"
   },
   "DataProcessing": {
      "InputFilter": "string",
      "JoinSource": "string",
      "OutputFilter": "string"
   },
   "Environment": {
      "string" : "string"
   },
   "ExperimentConfig": {
      "ExperimentName": "string",
      "RunName": "string",
      "TrialComponentDisplayName": "string",
      "TrialName": "string"
   },
   "FailureReason": "string",
   "LabelingJobArn": "string",
   "MaxConcurrentTransforms": number,
   "MaxPayloadInMB": number,
   "ModelClientConfig": {
      "InvocationsMaxRetries": number,
      "InvocationsTimeoutInSeconds": number
   },
   "ModelName": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "TransformEndTime": number,
   "TransformInput": {
      "CompressionType": "string",
      "ContentType": "string",
      "DataSource": {
         "S3DataSource": {
            "S3DataType": "string",
            "S3Uri": "string"
         }
```

```
            },
            "SplitType": "string"
         },
         "TransformJobArn": "string",
         "TransformJobName": "string",
         "TransformJobStatus": "string",
         "TransformOutput": {
            "Accept": "string",
            "AssembleWith": "string",
            "KmsKeyId": "string",
            "S3OutputPath": "string"
         },
         "TransformResources": {
            "InstanceCount": number,
            "InstanceType": "string",
            "TransformAmiVersion": "string",
            "VolumeKmsKeyId": "string"
         },
         "TransformStartTime": number
      },
      "Model": {
         "Containers": [
            {
               "AdditionalModelDataSources": [
                  {
                     "ChannelName": "string",
                     "S3DataSource": {
                        "CompressionType": "string",
                        "ETag": "string",
                        "HubAccessConfig": {
                           "HubContentArn": "string"
                        },
                        "ManifestEtag": "string",
                        "ManifestS3Uri": "string",
                        "ModelAccessConfig": {
                           "AcceptEula": boolean
                        },
                        "S3DataType": "string",
                        "S3Uri": "string"
                     }
                  }
               ],
               "ContainerHostname": "string",
               "ContainerMetricsConfig": {
```

```
                  "MetricsEndpoints": [
                     {
                        "MetricPublishFrequencyInSeconds": number,
                        "MetricsEndpointPath": "string"
                     }
                  ]
               },
               "Environment": {
                  "string" : "string"
               },
               "Image": "string",
               "ImageConfig": {
                  "RepositoryAccessMode": "string",
                  "RepositoryAuthConfig": {
                     "RepositoryCredentialsProviderArn": "string"
                  }
               },
               "InferenceSpecificationName": "string",
               "Mode": "string",
               "ModelDataSource": {
                  "S3DataSource": {
                     "CompressionType": "string",
                     "ETag": "string",
                     "HubAccessConfig": {
                        "HubContentArn": "string"
                     },
                     "ManifestEtag": "string",
                     "ManifestS3Uri": "string",
                     "ModelAccessConfig": {
                        "AcceptEula": boolean
                     },
                     "S3DataType": "string",
                     "S3Uri": "string"
                  }
               },
               "ModelDataUrl": "string",
               "ModelPackageName": "string",
               "MultiModelConfig": {
                  "ModelCacheSetting": "string"
               }
            }
         ],
         "CreationTime": number,
         "DeploymentRecommendation": {
```

```
      "RealTimeInferenceRecommendations": [
         {
            "Environment": {
               "string" : "string"
            },
            "InstanceType": "string",
            "RecommendationId": "string"
         }
      ],
      "RecommendationStatus": "string"
   },
   "EnableNetworkIsolation": boolean,
   "ExecutionRoleArn": "string",
   "InferenceExecutionConfig": {
      "Mode": "string"
   },
   "ModelArn": "string",
   "ModelName": "string",
   "PrimaryContainer": {
      "AdditionalModelDataSources": [
         {
            "ChannelName": "string",
            "S3DataSource": {
               "CompressionType": "string",
               "ETag": "string",
               "HubAccessConfig": {
                  "HubContentArn": "string"
               },
               "ManifestEtag": "string",
               "ManifestS3Uri": "string",
               "ModelAccessConfig": {
                  "AcceptEula": boolean
               },
               "S3DataType": "string",
               "S3Uri": "string"
            }
         }
      ],
      "ContainerHostname": "string",
      "ContainerMetricsConfig": {
         "MetricsEndpoints": [
            {
               "MetricPublishFrequencyInSeconds": number,
               "MetricsEndpointPath": "string"
```

```
                        }
                     ]
                  },
                  "Environment": {
                     "string" : "string"
                  },
                  "Image": "string",
                  "ImageConfig": {
                     "RepositoryAccessMode": "string",
                     "RepositoryAuthConfig": {
                        "RepositoryCredentialsProviderArn": "string"
                     }
                  },
                  "InferenceSpecificationName": "string",
                  "Mode": "string",
                  "ModelDataSource": {
                     "S3DataSource": {
                        "CompressionType": "string",
                        "ETag": "string",
                        "HubAccessConfig": {
                           "HubContentArn": "string"
                        },
                        "ManifestEtag": "string",
                        "ManifestS3Uri": "string",
                        "ModelAccessConfig": {
                           "AcceptEula": boolean
                        },
                        "S3DataType": "string",
                        "S3Uri": "string"
                     }
                  },
                  "ModelDataUrl": "string",
                  "ModelPackageName": "string",
                  "MultiModelConfig": {
                     "ModelCacheSetting": "string"
                  }
               },
               "Tags": [
                  {
                     "Key": "string",
                     "Value": "string"
                  }
               ],
               "VpcConfig": {
```

```
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         }
      },
      "ModelCard": {
         "CreatedBy": {
            "DomainId": "string",
            "IamIdentity": {
               "Arn": "string",
               "PrincipalId": "string",
               "SourceIdentity": "string"
            },
            "UserProfileArn": "string",
            "UserProfileName": "string"
         },
         "CreationTime": number,
         "LastModifiedBy": {
            "DomainId": "string",
            "IamIdentity": {
               "Arn": "string",
               "PrincipalId": "string",
               "SourceIdentity": "string"
            },
            "UserProfileArn": "string",
            "UserProfileName": "string"
         },
         "LastModifiedTime": number,
         "ModelCardArn": "string",
         "ModelCardName": "string",
         "ModelCardStatus": "string",
         "ModelCardVersion": number,
         "ModelId": "string",
         "RiskRating": "string",
         "SecurityConfig": {
            "KmsKeyId": "string"
         },
         "Tags": [
            {
               "Key": "string",
               "Value": "string"
            }
         ]
      },
      "MonitoringSchedules": [
```

```
            {
               "BatchTransformInput": {
                  "DataCapturedDestinationS3Uri": "string",
                  "DatasetFormat": {
                     "Csv": {
                        "Header": boolean
                     },
                     "Json": {
                        "Line": boolean
                     },
                     "Parquet": {
                     }
                  },
                  "EndTimeOffset": "string",
                  "ExcludeFeaturesAttribute": "string",
                  "FeaturesAttribute": "string",
                  "InferenceAttribute": "string",
                  "LocalPath": "string",
                  "ProbabilityAttribute": "string",
                  "ProbabilityThresholdAttribute": number,
                  "S3DataDistributionType": "string",
                  "S3InputMode": "string",
                  "StartTimeOffset": "string"
               },
               "CreationTime": number,
               "EndpointName": "string",
               "FailureReason": "string",
               "LastModifiedTime": number,
               "LastMonitoringExecutionSummary": {
                  "CreationTime": number,
                  "EndpointName": "string",
                  "FailureReason": "string",
                  "LastModifiedTime": number,
                  "MonitoringExecutionStatus": "string",
                  "MonitoringJobDefinitionName": "string",
                  "MonitoringScheduleName": "string",
                  "MonitoringType": "string",
                  "ProcessingJobArn": "string",
                  "ScheduledTime": number
               },
               "MonitoringAlertSummaries": [
                  {
                     "Actions": {
                        "ModelDashboardIndicator": {
```

```
                        "Enabled": boolean
                     }
                  },
                  "AlertStatus": "string",
                  "CreationTime": number,
                  "DatapointsToAlert": number,
                  "EvaluationPeriod": number,
                  "LastModifiedTime": number,
                  "MonitoringAlertName": "string"
               }
            ],
            "MonitoringScheduleArn": "string",
            "MonitoringScheduleConfig": {
               "MonitoringJobDefinition": {
                  "BaselineConfig": {
                     "BaseliningJobName": "string",
                     "ConstraintsResource": {
                        "S3Uri": "string"
                     },
                     "StatisticsResource": {
                        "S3Uri": "string"
                     }
                  },
                  "Environment": {
                     "string" : "string"
                  },
                  "MonitoringAppSpecification": {
                     "ContainerArguments": [ "string" ],
                     "ContainerEntrypoint": [ "string" ],
                     "ImageUri": "string",
                     "PostAnalyticsProcessorSourceUri": "string",
                     "RecordPreprocessorSourceUri": "string"
                  },
                  "MonitoringInputs": [
                     {
                        "BatchTransformInput": {
                           "DataCapturedDestinationS3Uri": "string",
                           "DatasetFormat": {
                              "Csv": {
                                 "Header": boolean
                              },
                              "Json": {
                                 "Line": boolean
                              },
```

```
                        "Parquet": {
                        }
                    },
                    "EndTimeOffset": "string",
                    "ExcludeFeaturesAttribute": "string",
                    "FeaturesAttribute": "string",
                    "InferenceAttribute": "string",
                    "LocalPath": "string",
                    "ProbabilityAttribute": "string",
                    "ProbabilityThresholdAttribute": number,
                    "S3DataDistributionType": "string",
                    "S3InputMode": "string",
                    "StartTimeOffset": "string"
                },
                "EndpointInput": {
                    "EndpointName": "string",
                    "EndTimeOffset": "string",
                    "ExcludeFeaturesAttribute": "string",
                    "FeaturesAttribute": "string",
                    "InferenceAttribute": "string",
                    "LocalPath": "string",
                    "ProbabilityAttribute": "string",
                    "ProbabilityThresholdAttribute": number,
                    "S3DataDistributionType": "string",
                    "S3InputMode": "string",
                    "StartTimeOffset": "string"
                }
            }
        ],
        "MonitoringOutputConfig": {
            "KmsKeyId": "string",
            "MonitoringOutputs": [
                {
                    "S3Output": {
                        "LocalPath": "string",
                        "S3UploadMode": "string",
                        "S3Uri": "string"
                    }
                }
            ]
        },
        "MonitoringResources": {
            "ClusterConfig": {
                "InstanceCount": number,
```

```
                        "InstanceType": "string",
                        "VolumeKmsKeyId": "string",
                        "VolumeSizeInGB": number
                     }
                  },
                  "NetworkConfig": {
                     "EnableInterContainerTrafficEncryption": boolean,
                     "EnableNetworkIsolation": boolean,
                     "VpcConfig": {
                        "SecurityGroupIds": [ "string" ],
                        "Subnets": [ "string" ]
                     }
                  },
                  "RoleArn": "string",
                  "StoppingCondition": {
                     "MaxRuntimeInSeconds": number
                  }
               },
               "MonitoringJobDefinitionName": "string",
               "MonitoringType": "string",
               "ScheduleConfig": {
                  "DataAnalysisEndTime": "string",
                  "DataAnalysisStartTime": "string",
                  "ScheduleExpression": "string"
               }
            },
            "MonitoringScheduleName": "string",
            "MonitoringScheduleStatus": "string",
            "MonitoringType": "string"
         }
      ]
   },
   "ModelCard": {
      "Content": "string",
      "CreatedBy": {
         "DomainId": "string",
         "IamIdentity": {
            "Arn": "string",
            "PrincipalId": "string",
            "SourceIdentity": "string"
         },
         "UserProfileArn": "string",
         "UserProfileName": "string"
      },
```

```
            "CreationTime": number,
            "LastModifiedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "LastModifiedTime": number,
            "ModelCardArn": "string",
            "ModelCardName": "string",
            "ModelCardStatus": "string",
            "ModelCardVersion": number,
            "ModelId": "string",
            "ModelPackageGroupName": "string",
            "RiskRating": "string",
            "SecurityConfig": {
               "KmsKeyId": "string"
            },
            "Tags": [
               {
                  "Key": "string",
                  "Value": "string"
               }
            ]
         },
         "ModelPackage": {
            "AdditionalInferenceSpecifications": [
               {
                  "Containers": [
                     {
                        "AdditionalModelDataSources": [
                           {
                              "ChannelName": "string",
                              "S3DataSource": {
                                 "CompressionType": "string",
                                 "ETag": "string",
                                 "HubAccessConfig": {
                                    "HubContentArn": "string"
                                 },
                                 "ManifestEtag": "string",
```

```
                "ManifestS3Uri": "string",
                "ModelAccessConfig": {
                   "AcceptEula": boolean
                },
                "S3DataType": "string",
                "S3Uri": "string"
             }
          }
       ],
       "AdditionalS3DataSource": {
          "CompressionType": "string",
          "ETag": "string",
          "S3DataType": "string",
          "S3Uri": "string"
       },
       "BaseModel": {
          "HubContentName": "string",
          "HubContentVersion": "string",
          "RecipeName": "string"
       },
       "ContainerHostname": "string",
       "Environment": {
          "string" : "string"
       },
       "Framework": "string",
       "FrameworkVersion": "string",
       "Image": "string",
       "ImageDigest": "string",
       "IsCheckpoint": boolean,
       "ModelDataETag": "string",
       "ModelDataSource": {
          "S3DataSource": {
             "CompressionType": "string",
             "ETag": "string",
             "HubAccessConfig": {
                "HubContentArn": "string"
             },
             "ManifestEtag": "string",
             "ManifestS3Uri": "string",
             "ModelAccessConfig": {
                "AcceptEula": boolean
             },
             "S3DataType": "string",
             "S3Uri": "string"
```

```
                        }
                    },
                    "ModelDataUrl": "string",
                    "ModelInput": {
                       "DataInputConfig": "string"
                    },
                    "NearestModelName": "string",
                    "ProductId": "string"
                 }
              ],
              "Description": "string",
              "Name": "string",
              "SupportedContentTypes": [ "string" ],
              "SupportedRealtimeInferenceInstanceTypes": [ "string" ],
              "SupportedResponseMIMETypes": [ "string" ],
              "SupportedTransformInstanceTypes": [ "string" ]
           }
        ],
        "ApprovalDescription": "string",
        "CertifyForMarketplace": boolean,
        "CreatedBy": {
           "DomainId": "string",
           "IamIdentity": {
              "Arn": "string",
              "PrincipalId": "string",
              "SourceIdentity": "string"
           },
           "UserProfileArn": "string",
           "UserProfileName": "string"
        },
        "CreationTime": number,
        "CustomerMetadataProperties": {
           "string" : "string"
        },
        "Domain": "string",
        "DriftCheckBaselines": {
           "Bias": {
              "ConfigFile": {
                 "ContentDigest": "string",
                 "ContentType": "string",
                 "S3Uri": "string"
              },
              "PostTrainingConstraints": {
                 "ContentDigest": "string",
```

```
            "ContentType": "string",
            "S3Uri": "string"
         },
         "PreTrainingConstraints": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
      },
      "Explainability": {
         "ConfigFile": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         },
         "Constraints": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
      },
      "ModelDataQuality": {
         "Constraints": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         },
         "Statistics": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
      },
      "ModelQuality": {
         "Constraints": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         },
         "Statistics": {
            "ContentDigest": "string",
            "ContentType": "string",
            "S3Uri": "string"
         }
```

```
            }
         },
         "InferenceSpecification": {
            "Containers": [
               {
                  "AdditionalModelDataSources": [
                     {
                        "ChannelName": "string",
                        "S3DataSource": {
                           "CompressionType": "string",
                           "ETag": "string",
                           "HubAccessConfig": {
                              "HubContentArn": "string"
                           },
                           "ManifestEtag": "string",
                           "ManifestS3Uri": "string",
                           "ModelAccessConfig": {
                              "AcceptEula": boolean
                           },
                           "S3DataType": "string",
                           "S3Uri": "string"
                        }
                     }
                  ],
                  "AdditionalS3DataSource": {
                     "CompressionType": "string",
                     "ETag": "string",
                     "S3DataType": "string",
                     "S3Uri": "string"
                  },
                  "BaseModel": {
                     "HubContentName": "string",
                     "HubContentVersion": "string",
                     "RecipeName": "string"
                  },
                  "ContainerHostname": "string",
                  "Environment": {
                     "string" : "string"
                  },
                  "Framework": "string",
                  "FrameworkVersion": "string",
                  "Image": "string",
                  "ImageDigest": "string",
                  "IsCheckpoint": boolean,
```

```
                "ModelDataETag": "string",
                "ModelDataSource": {
                   "S3DataSource": {
                      "CompressionType": "string",
                      "ETag": "string",
                      "HubAccessConfig": {
                         "HubContentArn": "string"
                      },
                      "ManifestEtag": "string",
                      "ManifestS3Uri": "string",
                      "ModelAccessConfig": {
                         "AcceptEula": boolean
                      },
                      "S3DataType": "string",
                      "S3Uri": "string"
                   }
                },
                "ModelDataUrl": "string",
                "ModelInput": {
                   "DataInputConfig": "string"
                },
                "NearestModelName": "string",
                "ProductId": "string"
             }
          ],
          "SupportedContentTypes": [ "string" ],
          "SupportedRealtimeInferenceInstanceTypes": [ "string" ],
          "SupportedResponseMIMETypes": [ "string" ],
          "SupportedTransformInstanceTypes": [ "string" ]
       },
       "LastModifiedBy": {
          "DomainId": "string",
          "IamIdentity": {
             "Arn": "string",
             "PrincipalId": "string",
             "SourceIdentity": "string"
          },
          "UserProfileArn": "string",
          "UserProfileName": "string"
       },
       "LastModifiedTime": number,
       "MetadataProperties": {
          "CommitId": "string",
          "GeneratedBy": "string",
```

```
            "ProjectId": "string",
            "Repository": "string"
         },
         "ModelApprovalStatus": "string",
         "ModelCard": {
            "ModelCardContent": "string",
            "ModelCardStatus": "string"
         },
         "ModelLifeCycle": {
            "Stage": "string",
            "StageDescription": "string",
            "StageStatus": "string"
         },
         "ModelMetrics": {
            "Bias": {
               "PostTrainingReport": {
                  "ContentDigest": "string",
                  "ContentType": "string",
                  "S3Uri": "string"
               },
               "PreTrainingReport": {
                  "ContentDigest": "string",
                  "ContentType": "string",
                  "S3Uri": "string"
               },
               "Report": {
                  "ContentDigest": "string",
                  "ContentType": "string",
                  "S3Uri": "string"
               }
            },
            "Explainability": {
               "Report": {
                  "ContentDigest": "string",
                  "ContentType": "string",
                  "S3Uri": "string"
               }
            },
            "ModelDataQuality": {
               "Constraints": {
                  "ContentDigest": "string",
                  "ContentType": "string",
                  "S3Uri": "string"
               },
```

```
            "Statistics": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            }
         },
         "ModelQuality": {
            "Constraints": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            },
            "Statistics": {
               "ContentDigest": "string",
               "ContentType": "string",
               "S3Uri": "string"
            }
         }
      },
      "ModelPackageArn": "string",
      "ModelPackageDescription": "string",
      "ModelPackageGroupName": "string",
      "ModelPackageName": "string",
      "ModelPackageRegistrationType": "string",
      "ModelPackageStatus": "string",
      "ModelPackageStatusDetails": {
         "ImageScanStatuses": [
            {
               "FailureReason": "string",
               "Name": "string",
               "Status": "string"
            }
         ],
         "ValidationStatuses": [
            {
               "FailureReason": "string",
               "Name": "string",
               "Status": "string"
            }
         ]
      },
      "ModelPackageVersion": number,
      "SamplePayloadUrl": "string",
      "SecurityConfig": {
```

```
      "KmsKeyId": "string"
   },
   "SkipModelValidation": "string",
   "SourceAlgorithmSpecification": {
      "SourceAlgorithms": [
         {
            "AlgorithmName": "string",
            "ModelDataETag": "string",
            "ModelDataSource": {
               "S3DataSource": {
                  "CompressionType": "string",
                  "ETag": "string",
                  "HubAccessConfig": {
                     "HubContentArn": "string"
                  },
                  "ManifestEtag": "string",
                  "ManifestS3Uri": "string",
                  "ModelAccessConfig": {
                     "AcceptEula": boolean
                  },
                  "S3DataType": "string",
                  "S3Uri": "string"
               }
            },
            "ModelDataUrl": "string"
         }
      ]
   },
   "SourceUri": "string",
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "Task": "string",
   "ValidationSpecification": {
      "ValidationProfiles": [
         {
            "ProfileName": "string",
            "TransformJobDefinition": {
               "BatchStrategy": "string",
               "Environment": {
                  "string" : "string"
```

```
                    },
                    "MaxConcurrentTransforms": number,
                    "MaxPayloadInMB": number,
                    "TransformInput": {
                       "CompressionType": "string",
                       "ContentType": "string",
                       "DataSource": {
                          "S3DataSource": {
                             "S3DataType": "string",
                             "S3Uri": "string"
                          }
                       },
                       "SplitType": "string"
                    },
                    "TransformOutput": {
                       "Accept": "string",
                       "AssembleWith": "string",
                       "KmsKeyId": "string",
                       "S3OutputPath": "string"
                    },
                    "TransformResources": {
                       "InstanceCount": number,
                       "InstanceType": "string",
                       "TransformAmiVersion": "string",
                       "VolumeKmsKeyId": "string"
                    }
                 }
              }
           ],
           "ValidationRole": "string"
        }
     },
     "ModelPackageGroup": {
        "CreatedBy": {
           "DomainId": "string",
           "IamIdentity": {
              "Arn": "string",
              "PrincipalId": "string",
              "SourceIdentity": "string"
           },
           "UserProfileArn": "string",
           "UserProfileName": "string"
        },
        "CreationTime": number,
```

```
            "ModelPackageGroupArn": "string",
            "ModelPackageGroupDescription": "string",
            "ModelPackageGroupName": "string",
            "ModelPackageGroupStatus": "string",
            "Tags": [
               {
                  "Key": "string",
                  "Value": "string"
               }
            ]
         },
         "Pipeline": {
            "CreatedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "CreationTime": number,
            "LastModifiedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "LastModifiedTime": number,
            "LastRunTime": number,
            "ParallelismConfiguration": {
               "MaxParallelExecutionSteps": number
            },
            "PipelineArn": "string",
            "PipelineDescription": "string",
            "PipelineDisplayName": "string",
            "PipelineName": "string",
            "PipelineStatus": "string",
            "RoleArn": "string",
```

```
            "Tags": [
               {
                  "Key": "string",
                  "Value": "string"
               }
            ]
         },
         "PipelineExecution": {
            "CreatedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "CreationTime": number,
            "FailureReason": "string",
            "LastModifiedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "LastModifiedTime": number,
            "ParallelismConfiguration": {
               "MaxParallelExecutionSteps": number
            },
            "PipelineArn": "string",
            "PipelineExecutionArn": "string",
            "PipelineExecutionDescription": "string",
            "PipelineExecutionDisplayName": "string",
            "PipelineExecutionStatus": "string",
            "PipelineExperimentConfig": {
               "ExperimentName": "string",
               "TrialName": "string"
            },
            "PipelineParameters": [
```

```
            {
               "Name": "string",
               "Value": "string"
            }
         ],
         "PipelineVersionDisplayName": "string",
         "PipelineVersionId": number,
         "SelectiveExecutionConfig": {
            "SelectedSteps": [
               {
                  "StepName": "string"
               }
            ],
            "SourcePipelineExecutionArn": "string"
         }
      },
      "PipelineVersion": {
         "CreatedBy": {
            "DomainId": "string",
            "IamIdentity": {
               "Arn": "string",
               "PrincipalId": "string",
               "SourceIdentity": "string"
            },
            "UserProfileArn": "string",
            "UserProfileName": "string"
         },
         "CreationTime": number,
         "LastExecutedPipelineExecutionArn": "string",
         "LastExecutedPipelineExecutionDisplayName": "string",
         "LastExecutedPipelineExecutionStatus": "string",
         "LastModifiedBy": {
            "DomainId": "string",
            "IamIdentity": {
               "Arn": "string",
               "PrincipalId": "string",
               "SourceIdentity": "string"
            },
            "UserProfileArn": "string",
            "UserProfileName": "string"
         },
         "LastModifiedTime": number,
         "PipelineArn": "string",
         "PipelineVersionDescription": "string",
```

```
            "PipelineVersionDisplayName": "string",
            "PipelineVersionId": number
         },
         "Project": {
            "CreatedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "CreationTime": number,
            "LastModifiedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "LastModifiedTime": number,
            "ProjectArn": "string",
            "ProjectDescription": "string",
            "ProjectId": "string",
            "ProjectName": "string",
            "ProjectStatus": "string",
            "ServiceCatalogProvisionedProductDetails": {
               "ProvisionedProductId": "string",
               "ProvisionedProductStatusMessage": "string"
            },
            "ServiceCatalogProvisioningDetails": {
               "PathId": "string",
               "ProductId": "string",
               "ProvisioningArtifactId": "string",
               "ProvisioningParameters": [
                  {
                     "Key": "string",
                     "Value": "string"
                  }
```

```
            ]
         },
         "Tags": [
            {
               "Key": "string",
               "Value": "string"
            }
         ],
         "TemplateProviderDetails": [
            {
               "CfnTemplateProviderDetail": {
                  "Parameters": [
                     {
                        "Key": "string",
                        "Value": "string"
                     }
                  ],
                  "RoleARN": "string",
                  "StackDetail": {
                     "Id": "string",
                     "Name": "string",
                     "StatusMessage": "string"
                  },
                  "TemplateName": "string",
                  "TemplateURL": "string"
               }
            }
         ]
      },
      "TrainingJob": {
         "AlgorithmSpecification": {
            "AlgorithmName": "string",
            "ContainerArguments": [ "string" ],
            "ContainerEntrypoint": [ "string" ],
            "EnableSageMakerMetricsTimeSeries": boolean,
            "MetricDefinitions": [
               {
                  "Name": "string",
                  "Regex": "string"
               }
            ],
            "TrainingImage": "string",
            "TrainingImageConfig": {
               "TrainingRepositoryAccessMode": "string",
```

```
            "TrainingRepositoryAuthConfig": {
               "TrainingRepositoryCredentialsProviderArn": "string"
            }
         },
         "TrainingInputMode": "string"
      },
      "AutoMLJobArn": "string",
      "BillableTimeInSeconds": number,
      "CheckpointConfig": {
         "LocalPath": "string",
         "S3Uri": "string"
      },
      "CreationTime": number,
      "DebugHookConfig": {
         "CollectionConfigurations": [
            {
               "CollectionName": "string",
               "CollectionParameters": {
                  "string" : "string"
               }
            }
         ],
         "HookParameters": {
            "string" : "string"
         },
         "LocalPath": "string",
         "S3OutputPath": "string"
      },
      "DebugRuleConfigurations": [
         {
            "InstanceType": "string",
            "LocalPath": "string",
            "RuleConfigurationName": "string",
            "RuleEvaluatorImage": "string",
            "RuleParameters": {
               "string" : "string"
            },
            "S3OutputPath": "string",
            "VolumeSizeInGB": number
         }
      ],
      "DebugRuleEvaluationStatuses": [
         {
            "LastModifiedTime": number,
```

```
            "RuleConfigurationName": "string",
            "RuleEvaluationJobArn": "string",
            "RuleEvaluationStatus": "string",
            "StatusDetails": "string"
         }
      ],
      "EnableInterContainerTrafficEncryption": boolean,
      "EnableManagedSpotTraining": boolean,
      "EnableNetworkIsolation": boolean,
      "Environment": {
         "string" : "string"
      },
      "ExperimentConfig": {
         "ExperimentName": "string",
         "RunName": "string",
         "TrialComponentDisplayName": "string",
         "TrialName": "string"
      },
      "FailureReason": "string",
      "FinalMetricDataList": [
         {
            "MetricName": "string",
            "Timestamp": number,
            "Value": number
         }
      ],
      "HyperParameters": {
         "string" : "string"
      },
      "InputDataConfig": [
         {
            "ChannelName": "string",
            "CompressionType": "string",
            "ContentType": "string",
            "DataSource": {
               "DatasetSource": {
                  "DatasetArn": "string"
               },
               "FileSystemDataSource": {
                  "DirectoryPath": "string",
                  "FileSystemAccessMode": "string",
                  "FileSystemId": "string",
                  "FileSystemType": "string"
               },
```

```
                "S3DataSource": {
                   "AttributeNames": [ "string" ],
                   "HubAccessConfig": {
                      "HubContentArn": "string"
                   },
                   "InstanceGroupNames": [ "string" ],
                   "ModelAccessConfig": {
                      "AcceptEula": boolean
                   },
                   "S3DataDistributionType": "string",
                   "S3DataType": "string",
                   "S3Uri": "string"
                }
             },
             "InputMode": "string",
             "RecordWrapperType": "string",
             "ShuffleConfig": {
                "Seed": number
             }
          }
       ],
       "LabelingJobArn": "string",
       "LastModifiedTime": number,
       "ModelArtifacts": {
          "S3ModelArtifacts": "string"
       },
       "ModelPackageConfig": {
          "ModelPackageGroupArn": "string",
          "SourceModelPackageArn": "string"
       },
       "OutputDataConfig": {
          "CompressionType": "string",
          "KmsKeyId": "string",
          "S3OutputPath": "string"
       },
       "OutputModelPackageArn": "string",
       "ProfilerConfig": {
          "DisableProfiler": boolean,
          "ProfilingIntervalInMilliseconds": number,
          "ProfilingParameters": {
             "string" : "string"
          },
          "S3OutputPath": "string"
       },
```

```
      "ResourceConfig": {
         "InstanceCount": number,
         "InstanceGroups": [
            {
               "InstanceCount": number,
               "InstanceGroupName": "string",
               "InstanceType": "string"
            }
         ],
         "InstancePlacementConfig": {
            "EnableMultipleJobs": boolean,
            "PlacementSpecifications": [
               {
                  "InstanceCount": number,
                  "UltraServerId": "string"
               }
            ]
         },
         "InstanceType": "string",
         "KeepAlivePeriodInSeconds": number,
         "TrainingPlanArn": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
      },
      "RetryStrategy": {
         "MaximumRetryAttempts": number
      },
      "RoleArn": "string",
      "SecondaryStatus": "string",
      "SecondaryStatusTransitions": [
         {
            "EndTime": number,
            "StartTime": number,
            "Status": "string",
            "StatusMessage": "string"
         }
      ],
      "StoppingCondition": {
         "MaxPendingTimeInSeconds": number,
         "MaxRuntimeInSeconds": number,
         "MaxWaitTimeInSeconds": number
      },
      "Tags": [
         {
```

```
                "Key": "string",
                "Value": "string"
            }
         ],
         "TensorBoardOutputConfig": {
            "LocalPath": "string",
            "S3OutputPath": "string"
         },
         "TrainingEndTime": number,
         "TrainingJobArn": "string",
         "TrainingJobName": "string",
         "TrainingJobStatus": "string",
         "TrainingStartTime": number,
         "TrainingTimeInSeconds": number,
         "TuningJobArn": "string",
         "VpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         },
         "WarmPoolStatus": {
            "ResourceRetainedBillableTimeInSeconds": number,
            "ReusedByJob": "string",
            "Status": "string"
         }
      },
      "Trial": {
         "CreatedBy": {
            "DomainId": "string",
            "IamIdentity": {
               "Arn": "string",
               "PrincipalId": "string",
               "SourceIdentity": "string"
            },
            "UserProfileArn": "string",
            "UserProfileName": "string"
         },
         "CreationTime": number,
         "DisplayName": "string",
         "ExperimentName": "string",
         "LastModifiedBy": {
            "DomainId": "string",
            "IamIdentity": {
               "Arn": "string",
               "PrincipalId": "string",
```

```
            "SourceIdentity": "string"
         },
         "UserProfileArn": "string",
         "UserProfileName": "string"
      },
      "LastModifiedTime": number,
      "MetadataProperties": {
         "CommitId": "string",
         "GeneratedBy": "string",
         "ProjectId": "string",
         "Repository": "string"
      },
      "Source": {
         "SourceArn": "string",
         "SourceType": "string"
      },
      "Tags": [
         {
            "Key": "string",
            "Value": "string"
         }
      ],
      "TrialArn": "string",
      "TrialComponentSummaries": [
         {
            "CreatedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "CreationTime": number,
            "TrialComponentArn": "string",
            "TrialComponentName": "string",
            "TrialComponentSource": {
               "SourceArn": "string",
               "SourceType": "string"
            }
         }
      ],
```

```
            "TrialName": "string"
         },
         "TrialComponent": {
            "CreatedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "CreationTime": number,
            "DisplayName": "string",
            "EndTime": number,
            "InputArtifacts": {
               "string" : {
                  "MediaType": "string",
                  "Value": "string"
               }
            },
            "LastModifiedBy": {
               "DomainId": "string",
               "IamIdentity": {
                  "Arn": "string",
                  "PrincipalId": "string",
                  "SourceIdentity": "string"
               },
               "UserProfileArn": "string",
               "UserProfileName": "string"
            },
            "LastModifiedTime": number,
            "LineageGroupArn": "string",
            "MetadataProperties": {
               "CommitId": "string",
               "GeneratedBy": "string",
               "ProjectId": "string",
               "Repository": "string"
            },
            "Metrics": [
               {
                  "Avg": number,
                  "Count": number,
```

```
            "Last": number,
            "Max": number,
            "MetricName": "string",
            "Min": number,
            "SourceArn": "string",
            "StdDev": number,
            "TimeStamp": number
         }
      ],
      "OutputArtifacts": {
         "string" : {
            "MediaType": "string",
            "Value": "string"
         }
      },
      "Parameters": {
         "string" : {
            "NumberValue": number,
            "StringValue": "string"
         }
      },
      "Parents": [
         {
            "ExperimentName": "string",
            "TrialName": "string"
         }
      ],
      "RunName": "string",
      "Source": {
         "SourceArn": "string",
         "SourceType": "string"
      },
      "SourceDetail": {
         "ProcessingJob": {
            "AppSpecification": {
               "ContainerArguments": [ "string" ],
               "ContainerEntrypoint": [ "string" ],
               "ImageUri": "string"
            },
            "AutoMLJobArn": "string",
            "CreationTime": number,
            "Environment": {
               "string" : "string"
            },
```

```
   "ExitMessage": "string",
   "ExperimentConfig": {
      "ExperimentName": "string",
      "RunName": "string",
      "TrialComponentDisplayName": "string",
      "TrialName": "string"
   },
   "FailureReason": "string",
   "LastModifiedTime": number,
   "MonitoringScheduleArn": "string",
   "NetworkConfig": {
      "EnableInterContainerTrafficEncryption": boolean,
      "EnableNetworkIsolation": boolean,
      "VpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ]
      }
   },
   "ProcessingEndTime": number,
   "ProcessingInputs": [
      {
         "AppManaged": boolean,
         "DatasetDefinition": {
            "AthenaDatasetDefinition": {
               "Catalog": "string",
               "Database": "string",
               "KmsKeyId": "string",
               "OutputCompression": "string",
               "OutputFormat": "string",
               "OutputS3Uri": "string",
               "QueryString": "string",
               "WorkGroup": "string"
            },
            "DataDistributionType": "string",
            "InputMode": "string",
            "LocalPath": "string",
            "RedshiftDatasetDefinition": {
               "ClusterId": "string",
               "ClusterRoleArn": "string",
               "Database": "string",
               "DbUser": "string",
               "KmsKeyId": "string",
               "OutputCompression": "string",
               "OutputFormat": "string",
```

```
                        "OutputS3Uri": "string",
                        "QueryString": "string"
                     }
                  },
                  "InputName": "string",
                  "S3Input": {
                     "LocalPath": "string",
                     "S3CompressionType": "string",
                     "S3DataDistributionType": "string",
                     "S3DataType": "string",
                     "S3InputMode": "string",
                     "S3Uri": "string"
                  }
               }
            ],
            "ProcessingJobArn": "string",
            "ProcessingJobName": "string",
            "ProcessingJobStatus": "string",
            "ProcessingOutputConfig": {
               "KmsKeyId": "string",
               "Outputs": [
                  {
                     "AppManaged": boolean,
                     "FeatureStoreOutput": {
                        "FeatureGroupName": "string"
                     },
                     "OutputName": "string",
                     "S3Output": {
                        "LocalPath": "string",
                        "S3UploadMode": "string",
                        "S3Uri": "string"
                     }
                  }
               ]
            },
            "ProcessingResources": {
               "ClusterConfig": {
                  "InstanceCount": number,
                  "InstanceType": "string",
                  "VolumeKmsKeyId": "string",
                  "VolumeSizeInGB": number
               }
            },
            "ProcessingStartTime": number,
```

```
            "RoleArn": "string",
            "StoppingCondition": {
               "MaxRuntimeInSeconds": number
            },
            "Tags": [
               {
                  "Key": "string",
                  "Value": "string"
               }
            ],
            "TrainingJobArn": "string"
         },
         "SourceArn": "string",
         "TrainingJob": {
            "AlgorithmSpecification": {
               "AlgorithmName": "string",
               "ContainerArguments": [ "string" ],
               "ContainerEntrypoint": [ "string" ],
               "EnableSageMakerMetricsTimeSeries": boolean,
               "MetricDefinitions": [
                  {
                     "Name": "string",
                     "Regex": "string"
                  }
               ],
               "TrainingImage": "string",
               "TrainingImageConfig": {
                  "TrainingRepositoryAccessMode": "string",
                  "TrainingRepositoryAuthConfig": {
                     "TrainingRepositoryCredentialsProviderArn": "string"
                  }
               },
               "TrainingInputMode": "string"
            },
            "AutoMLJobArn": "string",
            "BillableTimeInSeconds": number,
            "CheckpointConfig": {
               "LocalPath": "string",
               "S3Uri": "string"
            },
            "CreationTime": number,
            "DebugHookConfig": {
               "CollectionConfigurations": [
                  {
```

```
                "CollectionName": "string",
                "CollectionParameters": {
                   "string" : "string"
                }
             }
          ],
          "HookParameters": {
             "string" : "string"
          },
          "LocalPath": "string",
          "S3OutputPath": "string"
       },
       "DebugRuleConfigurations": [
          {
             "InstanceType": "string",
             "LocalPath": "string",
             "RuleConfigurationName": "string",
             "RuleEvaluatorImage": "string",
             "RuleParameters": {
                "string" : "string"
             },
             "S3OutputPath": "string",
             "VolumeSizeInGB": number
          }
       ],
       "DebugRuleEvaluationStatuses": [
          {
             "LastModifiedTime": number,
             "RuleConfigurationName": "string",
             "RuleEvaluationJobArn": "string",
             "RuleEvaluationStatus": "string",
             "StatusDetails": "string"
          }
       ],
       "EnableInterContainerTrafficEncryption": boolean,
       "EnableManagedSpotTraining": boolean,
       "EnableNetworkIsolation": boolean,
       "Environment": {
          "string" : "string"
       },
       "ExperimentConfig": {
          "ExperimentName": "string",
          "RunName": "string",
          "TrialComponentDisplayName": "string",
```

```
            "TrialName": "string"
         },
         "FailureReason": "string",
         "FinalMetricDataList": [
            {
               "MetricName": "string",
               "Timestamp": number,
               "Value": number
            }
         ],
         "HyperParameters": {
            "string" : "string"
         },
         "InputDataConfig": [
            {
               "ChannelName": "string",
               "CompressionType": "string",
               "ContentType": "string",
               "DataSource": {
                  "DatasetSource": {
                     "DatasetArn": "string"
                  },
                  "FileSystemDataSource": {
                     "DirectoryPath": "string",
                     "FileSystemAccessMode": "string",
                     "FileSystemId": "string",
                     "FileSystemType": "string"
                  },
                  "S3DataSource": {
                     "AttributeNames": [ "string" ],
                     "HubAccessConfig": {
                        "HubContentArn": "string"
                     },
                     "InstanceGroupNames": [ "string" ],
                     "ModelAccessConfig": {
                        "AcceptEula": boolean
                     },
                     "S3DataDistributionType": "string",
                     "S3DataType": "string",
                     "S3Uri": "string"
                  }
               },
               "InputMode": "string",
               "RecordWrapperType": "string",
```

```
            "ShuffleConfig": {
               "Seed": number
            }
         }
      ],
      "LabelingJobArn": "string",
      "LastModifiedTime": number,
      "ModelArtifacts": {
         "S3ModelArtifacts": "string"
      },
      "ModelPackageConfig": {
         "ModelPackageGroupArn": "string",
         "SourceModelPackageArn": "string"
      },
      "OutputDataConfig": {
         "CompressionType": "string",
         "KmsKeyId": "string",
         "S3OutputPath": "string"
      },
      "OutputModelPackageArn": "string",
      "ProfilerConfig": {
         "DisableProfiler": boolean,
         "ProfilingIntervalInMilliseconds": number,
         "ProfilingParameters": {
            "string" : "string"
         },
         "S3OutputPath": "string"
      },
      "ResourceConfig": {
         "InstanceCount": number,
         "InstanceGroups": [
            {
               "InstanceCount": number,
               "InstanceGroupName": "string",
               "InstanceType": "string"
            }
         ],
         "InstancePlacementConfig": {
            "EnableMultipleJobs": boolean,
            "PlacementSpecifications": [
               {
                  "InstanceCount": number,
                  "UltraServerId": "string"
               }
```

```
            ]
         },
         "InstanceType": "string",
         "KeepAlivePeriodInSeconds": number,
         "TrainingPlanArn": "string",
         "VolumeKmsKeyId": "string",
         "VolumeSizeInGB": number
      },
      "RetryStrategy": {
         "MaximumRetryAttempts": number
      },
      "RoleArn": "string",
      "SecondaryStatus": "string",
      "SecondaryStatusTransitions": [
         {
            "EndTime": number,
            "StartTime": number,
            "Status": "string",
            "StatusMessage": "string"
         }
      ],
      "StoppingCondition": {
         "MaxPendingTimeInSeconds": number,
         "MaxRuntimeInSeconds": number,
         "MaxWaitTimeInSeconds": number
      },
      "Tags": [
         {
            "Key": "string",
            "Value": "string"
         }
      ],
      "TensorBoardOutputConfig": {
         "LocalPath": "string",
         "S3OutputPath": "string"
      },
      "TrainingEndTime": number,
      "TrainingJobArn": "string",
      "TrainingJobName": "string",
      "TrainingJobStatus": "string",
      "TrainingStartTime": number,
      "TrainingTimeInSeconds": number,
      "TuningJobArn": "string",
      "VpcConfig": {
```

```
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         },
         "WarmPoolStatus": {
            "ResourceRetainedBillableTimeInSeconds": number,
            "ReusedByJob": "string",
            "Status": "string"
         }
      },
      "TransformJob": {
         "AutoMLJobArn": "string",
         "BatchStrategy": "string",
         "CreationTime": number,
         "DataCaptureConfig": {
            "DestinationS3Uri": "string",
            "GenerateInferenceId": boolean,
            "KmsKeyId": "string"
         },
         "DataProcessing": {
            "InputFilter": "string",
            "JoinSource": "string",
            "OutputFilter": "string"
         },
         "Environment": {
            "string" : "string"
         },
         "ExperimentConfig": {
            "ExperimentName": "string",
            "RunName": "string",
            "TrialComponentDisplayName": "string",
            "TrialName": "string"
         },
         "FailureReason": "string",
         "LabelingJobArn": "string",
         "MaxConcurrentTransforms": number,
         "MaxPayloadInMB": number,
         "ModelClientConfig": {
            "InvocationsMaxRetries": number,
            "InvocationsTimeoutInSeconds": number
         },
         "ModelName": "string",
         "Tags": [
            {
               "Key": "string",
```

```
                        "Value": "string"
                    }
                ],
                "TransformEndTime": number,
                "TransformInput": {
                    "CompressionType": "string",
                    "ContentType": "string",
                    "DataSource": {
                        "S3DataSource": {
                            "S3DataType": "string",
                            "S3Uri": "string"
                        }
                    },
                    "SplitType": "string"
                },
                "TransformJobArn": "string",
                "TransformJobName": "string",
                "TransformJobStatus": "string",
                "TransformOutput": {
                    "Accept": "string",
                    "AssembleWith": "string",
                    "KmsKeyId": "string",
                    "S3OutputPath": "string"
                },
                "TransformResources": {
                    "InstanceCount": number,
                    "InstanceType": "string",
                    "TransformAmiVersion": "string",
                    "VolumeKmsKeyId": "string"
                },
                "TransformStartTime": number
            }
        },
        "StartTime": number,
        "Status": {
            "Message": "string",
            "PrimaryStatus": "string"
        },
        "Tags": [
            {
                "Key": "string",
                "Value": "string"
            }
        ],
```

```
            "TrialComponentArn": "string",
            "TrialComponentName": "string"
         }
      }
   ],
   "TotalHits": {
      "Relation": "string",
      "Value": number
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the result of the previous `Search` request was truncated, the response includes a NextToken. To retrieve the next set of results, use the token in the next request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Pattern: `.*`

Results

A list of `SearchRecord` objects.

Type: Array of SearchRecord objects

TotalHits

The total number of matching results.

Type: TotalHits object

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

SearchTrainingPlanOfferings

Service: Amazon SageMaker Service

Searches for available training plan offerings based on specified criteria.

- Users search for available plan offerings based on their requirements (e.g., instance type, count, start time, duration).
- And then, they create a plan that best matches their needs using the ID of the plan offering they want to use.

For more information about how to reserve GPU capacity for your SageMaker training jobs or SageMaker HyperPod clusters using Amazon SageMaker Training Plan , see `CreateTrainingPlan` .

Request Syntax

```
{
   "DurationHours": number,
   "EndTimeBefore": number,
   "InstanceCount": number,
   "InstanceType": "string",
   "StartTimeAfter": number,
   "TargetResources": [ "string" ],
   "TrainingPlanArn": "string",
   "UltraServerCount": number,
   "UltraServerType": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DurationHours

The desired duration in hours for the training plan offerings.

Type: Long

Valid Range: Minimum value of 1. Maximum value of 87600.

Required: No

EndTimeBefore

A filter to search for reserved capacity offerings with an end time before a specified date.

Type: Timestamp

Required: No

InstanceCount

The number of instances you want to reserve in the training plan offerings. This allows you to specify the quantity of compute resources needed for your SageMaker training jobs or SageMaker HyperPod clusters, helping you find reserved capacity offerings that match your requirements.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 256.

Required: No

InstanceType

The type of instance you want to search for in the available training plan offerings. This field allows you to filter the search results based on the specific compute resources you require for your SageMaker training jobs or SageMaker HyperPod clusters. When searching for training plan offerings, specifying the instance type helps you find Reserved Instances that match your computational needs.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.trn1.32xlarge | ml.trn2.48xlarge | ml.p6-b200.48xlarge | ml.p4de.24xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge | ml.p6-b300.48xlarge`

Required: No

StartTimeAfter

A filter to search for training plan offerings with a start time after a specified date.

Type: Timestamp

Required: No

TargetResources

The target resources (e.g., SageMaker Training Jobs, SageMaker HyperPod, SageMaker Endpoints, Studio apps) to search for in the offerings.

Training plans are specific to their target resource.

- A training plan designed for SageMaker training jobs can only be used to schedule and run training jobs.
- A training plan for HyperPod clusters can be used exclusively to provide compute resources to a cluster's instance group.
- A training plan for SageMaker endpoints can be used exclusively to provide compute resources to SageMaker endpoints for model deployment.
- A training plan for Studio apps can be used to launch JupyterLab and Code Editor apps on reserved training plan capacity.

Type: Array of strings

Array Members: Minimum number of 1 item.

Valid Values: `training-job | hyperpod-cluster | endpoint | studio-apps`

Required: No

TrainingPlanArn

The Amazon Resource Name (ARN); of an existing training plan to search for extension offerings. When specified, the API returns extension offerings that can be used to extend the specified training plan.

Type: String

Required: No

UltraServerCount

The number of UltraServers to search for.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

UltraServerType

The type of UltraServer to search for, such as ml.u-p6e-gb200x72.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `ml.[a-z0-9\-.]+`

Required: No

Response Syntax

```
{
   "TrainingPlanExtensionOfferings": [
      {
         "AvailabilityZone": "string",
         "CurrencyCode": "string",
         "DurationHours": number,
         "EndDate": number,
         "StartDate": number,
         "TrainingPlanExtensionOfferingId": "string",
         "UpfrontFee": "string"
      }
   ],
   "TrainingPlanOfferings": [
      {
         "CurrencyCode": "string",
         "DurationHours": number,
         "DurationMinutes": number,
         "RequestedEndTimeBefore": number,
         "RequestedStartTimeAfter": number,
         "ReservedCapacityOfferings": [
            {
               "AvailabilityZone": "string",
               "DurationHours": number,
               "DurationMinutes": number,
               "EndTime": number,
```

```
                "ExtensionEndTime": number,
                "ExtensionStartTime": number,
                "InstanceCount": number,
                "InstanceType": "string",
                "ReservedCapacityType": "string",
                "StartTime": number,
                "UltraServerCount": number,
                "UltraServerType": "string"
            }
        ],
        "TargetResources": [ "string" ],
        "TrainingPlanOfferingId": "string",
        "UpfrontFee": "string"
    }
  ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrainingPlanExtensionOfferings

A list of extension offerings available for the specified training plan. These offerings can be used with the `ExtendTrainingPlan` API to extend an existing training plan.

Type: Array of TrainingPlanExtensionOffering objects

Array Members: Minimum number of 0 items.

TrainingPlanOfferings

A list of training plan offerings that match the search criteria.

Type: Array of TrainingPlanOffering objects

Array Members: Minimum number of 0 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

SendPipelineExecutionStepFailure

Service: Amazon SageMaker Service

Notifies the pipeline that the execution of a callback step failed, along with a message describing why. When a callback step is run, the pipeline generates a callback token and includes the token in a message sent to Amazon Simple Queue Service (Amazon SQS).

Request Syntax

```
{
   "CallbackToken": "string",
   "ClientRequestToken": "string",
   "FailureReason": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CallbackToken

The pipeline generated token from the Amazon SQS queue.

Type: String

Length Constraints: Fixed length of 10.

Pattern: `[a-zA-Z0-9]+`

Required: Yes

ClientRequestToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the operation. An idempotent operation completes no more than one time.

Type: String

Length Constraints: Minimum length of 32. Maximum length of 128.

Required: No

FailureReason

A message describing why the step failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

Response Syntax

```
{
   "PipelineExecutionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

SendPipelineExecutionStepSuccess

Service: Amazon SageMaker Service

Notifies the pipeline that the execution of a callback step succeeded and provides a list of the step's output parameters. When a callback step is run, the pipeline generates a callback token and includes the token in a message sent to Amazon Simple Queue Service (Amazon SQS).

Request Syntax

```
{
   "CallbackToken": "string",
   "ClientRequestToken": "string",
   "OutputParameters": [
      {
         "Name": "string",
         "Value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CallbackToken

The pipeline generated token from the Amazon SQS queue.

Type: String

Length Constraints: Fixed length of 10.

Pattern: `[a-zA-Z0-9]+`

Required: Yes

ClientRequestToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the operation. An idempotent operation completes no more than one time.

Type: String

Length Constraints: Minimum length of 32. Maximum length of 128.

Required: No

OutputParameters

A list of the output parameters of the callback step.

Type: Array of OutputParameter objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Response Syntax

```
{
   "PipelineExecutionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StartClusterHealthCheck

Service: Amazon SageMaker Service

Start deep health checks for a SageMaker HyperPod cluster. You can use DescribeClusterNode API to track progress of the deep health checks. The unhealthy nodes will be automatically rebooted or replaced. Please see Resilience-related Kubernetes labels by SageMaker HyperPod for details.

Request Syntax

```
{
   "ClusterName": "string",
   "DeepHealthCheckConfigurations": [
      {
         "DeepHealthChecks": [ "string" ],
         "InstanceGroupName": "string",
         "InstanceIds": [ "string" ]
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterName

The string name or the Amazon Resource Name (ARN) of the SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

DeepHealthCheckConfigurations

A list of configurations containing instance group names, EC2 instance IDs, and deep health checks to perform.

Type: Array of InstanceGroupHealthCheckConfiguration objects

Array Members: Minimum number of 1 item. Maximum number of 99 items.

Required: Yes

Response Syntax

```
{
   "ClusterArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ClusterArn

The Amazon Resource Name (ARN) of the SageMaker HyperPod cluster on which the deep health checks were initiated.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StartEdgeDeploymentStage

Service: Amazon SageMaker Service

Starts a stage in an edge deployment plan.

Request Syntax

```
{
   "EdgeDeploymentPlanName": "string",
   "StageName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EdgeDeploymentPlanName

The name of the edge deployment plan to start.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

StageName

The name of the stage to start.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StartInferenceExperiment

Service: Amazon SageMaker Service

Starts an inference experiment.

Request Syntax

```
{
   "Name": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Name

The name of the inference experiment to start.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "InferenceExperimentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

InferenceExperimentArn

The ARN of the started inference experiment to start.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:inference-experiment/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StartMlflowTrackingServer

Service: Amazon SageMaker Service

Programmatically start an MLflow Tracking Server.

Request Syntax

```
{
   "TrackingServerName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrackingServerName

The name of the tracking server to start.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: Yes

Response Syntax

```
{
   "TrackingServerArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrackingServerArn

The ARN of the started tracking server.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-tracking-server/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StartMonitoringSchedule

Service: Amazon SageMaker Service

Starts a previously stopped monitoring schedule.

> **Note**
>
> By default, when you successfully create a new schedule, the status of a monitoring schedule is `scheduled`.

Request Syntax

```
{
   "MonitoringScheduleName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MonitoringScheduleName

The name of the schedule to start.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StartNotebookInstance

Service: Amazon SageMaker Service

Launches an ML compute instance with the latest version of the libraries and attaches your ML storage volume. After configuring the notebook instance, SageMaker AI sets the notebook instance status to `InService`. A notebook instance's status must be `InService` before you can connect to your Jupyter notebook.

Request Syntax

```
{
   "NotebookInstanceName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

NotebookInstanceName

The name of the notebook instance to start.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StartPipelineExecution

Service: Amazon SageMaker Service

Starts a pipeline execution.

Request Syntax

```
{
   "ClientRequestToken": "string",
   "MlflowExperimentName": "string",
   "ParallelismConfiguration": {
      "MaxParallelExecutionSteps": number
   },
   "PipelineExecutionDescription": "string",
   "PipelineExecutionDisplayName": "string",
   "PipelineName": "string",
   "PipelineParameters": [
      {
         "Name": "string",
         "Value": "string"
      }
   ],
   "PipelineVersionId": number,
   "SelectiveExecutionConfig": {
      "SelectedSteps": [
         {
            "StepName": "string"
         }
      ],
      "SourcePipelineExecutionArn": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClientRequestToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the operation. An idempotent operation completes no more than once.

Type: String

Length Constraints: Minimum length of 32. Maximum length of 128.

Required: Yes

MlflowExperimentName

The MLflow experiment name of the pipeline execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*`

Required: No

ParallelismConfiguration

This configuration, if specified, overrides the parallelism configuration of the parent pipeline for this specific run.

Type: ParallelismConfiguration object

Required: No

PipelineExecutionDescription

The description of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

PipelineExecutionDisplayName

The display name of the pipeline execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 82.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,81}`

Required: No

PipelineName

The name or Amazon Resource Name (ARN) of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,255})`

Required: Yes

PipelineParameters

Contains a list of pipeline parameters. This list can be empty.

Type: Array of Parameter objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

PipelineVersionId

The ID of the pipeline version to start execution from.

Type: Long

Valid Range: Minimum value of 1.

Required: No

SelectiveExecutionConfig

The selective execution configuration applied to the pipeline run.

Type: SelectiveExecutionConfig object

Required: No

Response Syntax

```
{
```

```
   "PipelineExecutionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StartSession

Service: Amazon SageMaker Service

Initiates a remote connection session between a local integrated development environments (IDEs) and a remote SageMaker space.

Request Syntax

```
{
   "ResourceIdentifier": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ResourceIdentifier

The Amazon Resource Name (ARN) of the resource to which the remote connection will be established. For example, this identifies the specific ARN space application you want to connect to from your local IDE.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:.*\/.*`

Required: Yes

Response Syntax

```
{
   "SessionId": "string",
   "StreamUrl": "string",
   "TokenValue": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

SessionId

A unique identifier for the established remote connection session.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

StreamUrl

A WebSocket URL used to establish a SSH connection between the local IDE and remote SageMaker space.

Type: String

TokenValue

An encrypted token value containing session and caller information.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopAIBenchmarkJob

Service: Amazon SageMaker Service

Stops a running AI benchmark job.

Request Syntax

```
{
   "AIBenchmarkJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AIBenchmarkJobName

The name of the AI benchmark job to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AIBenchmarkJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIBenchmarkJobArn

The Amazon Resource Name (ARN) of the stopped benchmark job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-benchmark-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopAIRecommendationJob

Service: Amazon SageMaker Service

Stops a running AI recommendation job.

Request Syntax

```
{
   "AIRecommendationJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AIRecommendationJobName

The name of the AI recommendation job to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "AIRecommendationJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AIRecommendationJobArn

The Amazon Resource Name (ARN) of the stopped recommendation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-recommendation-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopAutoMLJob

Service: Amazon SageMaker Service

A method for forcing a running job to shut down.

Request Syntax

```
{
   "AutoMLJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AutoMLJobName

The name of the object you are requesting.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopCompilationJob

Service: Amazon SageMaker Service

Stops a model compilation job.

To stop a job, Amazon SageMaker AI sends the algorithm the SIGTERM signal. This gracefully shuts the job down. If the job hasn't stopped, it sends the SIGKILL signal.

When it receives a `StopCompilationJob` request, Amazon SageMaker AI changes the `CompilationJobStatus` of the job to `Stopping`. After Amazon SageMaker stops the job, it sets the `CompilationJobStatus` to `Stopped`.

Request Syntax

```
{
   "CompilationJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CompilationJobName

The name of the model compilation job to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopEdgeDeploymentStage

Service: Amazon SageMaker Service

Stops a stage in an edge deployment plan.

Request Syntax

```
{
   "EdgeDeploymentPlanName": "string",
   "StageName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EdgeDeploymentPlanName

The name of the edge deployment plan to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

StageName

The name of the stage to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopEdgePackagingJob

Service: Amazon SageMaker Service

Request to stop an edge packaging job.

Request Syntax

```
{
   "EdgePackagingJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

EdgePackagingJobName

The name of the edge packaging job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopHyperParameterTuningJob

Service: Amazon SageMaker Service

Stops a running hyperparameter tuning job and all running training jobs that the tuning job launched.

All model artifacts output from the training jobs are stored in Amazon Simple Storage Service (Amazon S3). All data that the training jobs write to Amazon CloudWatch Logs are still available in CloudWatch. After the tuning job moves to the `Stopped` state, it releases all reserved resources for the tuning job.

Request Syntax

```
{
   "HyperParameterTuningJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HyperParameterTuningJobName

The name of the tuning job to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopInferenceExperiment

Service: Amazon SageMaker Service

Stops an inference experiment.

Request Syntax

```
{
   "DesiredModelVariants": [
      {
         "InfrastructureConfig": {
            "InfrastructureType": "string",
            "RealTimeInferenceConfig": {
               "InstanceCount": number,
               "InstanceType": "string"
            }
         },
         "ModelName": "string",
         "VariantName": "string"
      }
   ],
   "DesiredState": "string",
   "ModelVariantActions": {
      "string" : "string"
   },
   "Name": "string",
   "Reason": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DesiredModelVariants

An array of `ModelVariantConfig` objects. There is one for each variant that you want to deploy after the inference experiment stops. Each `ModelVariantConfig` describes the infrastructure configuration for deploying the corresponding variant.

Type: Array of ModelVariantConfig objects

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Required: No

DesiredState

The desired state of the experiment after stopping. The possible states are the following:

- `Completed`: The experiment completed successfully
- `Cancelled`: The experiment was canceled

Type: String

Valid Values: `Completed | Cancelled`

Required: No

ModelVariantActions

Array of key-value pairs, with names of variants mapped to actions. The possible actions are the following:

- `Promote` - Promote the shadow variant to a production variant
- `Remove` - Delete the variant
- `Retain` - Keep the variant as it is

Type: String to string map

Map Entries: Maximum number of 2 items.

Key Length Constraints: Minimum length of 0. Maximum length of 63.

Key Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Valid Values: `Retain | Remove | Promote`

Required: Yes

Name

The name of the inference experiment to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Reason

The reason for stopping the experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "InferenceExperimentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

InferenceExperimentArn

The ARN of the stopped inference experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:inference-experiment/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopInferenceRecommendationsJob

Service: Amazon SageMaker Service

Stops an Inference Recommender job.

Request Syntax

```
{
   "JobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobName

The name of the job you want to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopJob

Service: Amazon SageMaker Service

Stops a running job. When you call `StopJob`, Amazon SageMaker sets the job status to `Stopping`. After the job stops, the status changes to `Stopped`. Partial results may be available in the output location if the job was in progress. To delete a stopped job, call `DeleteJob`.

The following operations are related to `StopJob`:

- `CreateJob`
- `DescribeJob`
- `DeleteJob`

Request Syntax

```
{
   "JobCategory": "string",
   "JobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

JobCategory

The category of the job to stop.

Type: String

Valid Values: `AgentRFT | AgentRFTEvaluation`

Required: Yes

JobName

The name of the job to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopLabelingJob

Service: Amazon SageMaker Service

Stops a running labeling job. A job that is stopped cannot be restarted. Any results obtained before the job is stopped are placed in the Amazon S3 output bucket.

Request Syntax

```
{
   "LabelingJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

LabelingJobName

The name of the labeling job to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopMlflowTrackingServer

Service: Amazon SageMaker Service

Programmatically stop an MLflow Tracking Server.

Request Syntax

```
{
   "TrackingServerName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrackingServerName

The name of the tracking server to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: Yes

Response Syntax

```
{
   "TrackingServerArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrackingServerArn

The ARN of the stopped tracking server.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-tracking-server/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopMonitoringSchedule

Service: Amazon SageMaker Service

Stops a previously started monitoring schedule.

Request Syntax

```
{
   "MonitoringScheduleName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MonitoringScheduleName

The name of the schedule to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopNotebookInstance

Service: Amazon SageMaker Service

Terminates the ML compute instance. Before terminating the instance, SageMaker AI disconnects the ML storage volume from it. SageMaker AI preserves the ML storage volume. SageMaker AI stops charging you for the ML compute instance when you call `StopNotebookInstance`.

To access data on the ML storage volume for a notebook instance that has been terminated, call the `StartNotebookInstance` API. `StartNotebookInstance` launches another ML compute instance, configures it, and attaches the preserved ML storage volume so you can continue your work.

Request Syntax

```
{
   "NotebookInstanceName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

NotebookInstanceName

The name of the notebook instance to terminate.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopOptimizationJob

Service: Amazon SageMaker Service

Ends a running inference optimization job.

Request Syntax

```
{
   "OptimizationJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

OptimizationJobName

The name that you assigned to the optimization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopPipelineExecution

Service: Amazon SageMaker Service

Stops a pipeline execution.

Callback Step

A pipeline execution won't stop while a callback step is running. When you call `StopPipelineExecution` on a pipeline execution with a running callback step, SageMaker Pipelines sends an additional Amazon SQS message to the specified SQS queue. The body of the SQS message contains a "Status" field which is set to "Stopping".

You should add logic to your Amazon SQS message consumer to take any needed action (for example, resource cleanup) upon receipt of the message followed by a call to `SendPipelineExecutionStepSuccess` or `SendPipelineExecutionStepFailure`.

Only when SageMaker Pipelines receives one of these calls will it stop the pipeline execution.

Lambda Step

A pipeline execution can't be stopped while a lambda step is running because the Lambda function invoked by the lambda step can't be stopped. If you attempt to stop the execution while the Lambda function is running, the pipeline waits for the Lambda function to finish or until the timeout is hit, whichever occurs first, and then stops. If the Lambda function finishes, the pipeline execution status is `Stopped`. If the timeout is hit the pipeline execution status is `Failed`.

Request Syntax

```
{
   "ClientRequestToken": "string",
   "PipelineExecutionArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClientRequestToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the operation. An idempotent operation completes no more than once.

Type: String

Length Constraints: Minimum length of 32. Maximum length of 128.

Required: Yes

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: Yes

Response Syntax

```
{
   "PipelineExecutionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopProcessingJob

Service: Amazon SageMaker Service

Stops a processing job.

Request Syntax

```
{
   "ProcessingJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ProcessingJobName

The name of the processing job to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopTrainingJob

Service: Amazon SageMaker Service

Stops a training job. To stop a job, SageMaker sends the algorithm the `SIGTERM` signal, which delays job termination for 120 seconds. Algorithms might use this 120-second window to save the model artifacts, so the results of the training is not lost.

When it receives a `StopTrainingJob` request, SageMaker changes the status of the job to `Stopping`. After SageMaker stops the job, it sets the status to `Stopped`.

Request Syntax

```
{
   "TrainingJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TrainingJobName

The name of the training job to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopTransformJob

Service: Amazon SageMaker Service

Stops a batch transform job.

When Amazon SageMaker receives a `StopTransformJob` request, the status of the job changes to `Stopping`. After Amazon SageMaker stops the job, the status is set to `Stopped`. When you stop a batch transform job before it is completed, Amazon SageMaker doesn't store the job's output in Amazon S3.

Request Syntax

```
{
   "TransformJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

TransformJobName

The name of the batch transform job to stop.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateAction

Service: Amazon SageMaker Service

Updates an action.

Request Syntax

```
{
   "ActionName": "string",
   "Description": "string",
   "Properties": {
      "string" : "string"
   },
   "PropertiesToRemove": [ "string" ],
   "Status": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ActionName

The name of the action to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Description

The new description for the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

Properties

The new list of properties. Overwrites the current property list.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 30 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2500.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Required: No

PropertiesToRemove

A list of properties to remove.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 2500.

Pattern: `.*`

Required: No

Status

The new status for the action.

Type: String

Valid Values: `Unknown | InProgress | Completed | Failed | Stopping | Stopped`

Required: No

Response Syntax

```
{
```

```
   "ActionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ActionArn

The Amazon Resource Name (ARN) of the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:action/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateAppImageConfig

Service: Amazon SageMaker Service

Updates the properties of an AppImageConfig.

Request Syntax

```
{
   "AppImageConfigName": "string",
   "CodeEditorAppImageConfig": {
      "ContainerConfig": {
         "ContainerArguments": [ "string" ],
         "ContainerEntrypoint": [ "string" ],
         "ContainerEnvironmentVariables": {
            "string" : "string"
         }
      },
      "FileSystemConfig": {
         "DefaultGid": number,
         "DefaultUid": number,
         "MountPath": "string"
      }
   },
   "JupyterLabAppImageConfig": {
      "ContainerConfig": {
         "ContainerArguments": [ "string" ],
         "ContainerEntrypoint": [ "string" ],
         "ContainerEnvironmentVariables": {
            "string" : "string"
         }
      },
      "FileSystemConfig": {
         "DefaultGid": number,
         "DefaultUid": number,
         "MountPath": "string"
      }
   },
   "KernelGatewayImageConfig": {
      "FileSystemConfig": {
         "DefaultGid": number,
         "DefaultUid": number,
         "MountPath": "string"
      },
```

```
      "KernelSpecs": [
         {
            "DisplayName": "string",
            "Name": "string"
         }
      ]
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AppImageConfigName

The name of the AppImageConfig to update.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

CodeEditorAppImageConfig

The Code Editor app running on the image.

Type: CodeEditorAppImageConfig object

Required: No

JupyterLabAppImageConfig

The JupyterLab app running on the image.

Type: JupyterLabAppImageConfig object

Required: No

KernelGatewayImageConfig

The new KernelGateway app to run on the image.

Type: KernelGatewayImageConfig object

Required: No

Response Syntax

```
{
   "AppImageConfigArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

AppImageConfigArn

The ARN for the AppImageConfig.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:app-image-config/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateArtifact

Service: Amazon SageMaker Service

Updates an artifact.

Request Syntax

```
{
   "ArtifactArn": "string",
   "ArtifactName": "string",
   "Properties": {
      "string" : "string"
   },
   "PropertiesToRemove": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ArtifactArn

The Amazon Resource Name (ARN) of the artifact to update.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:artifact/.*`

Required: Yes

ArtifactName

The new name for the artifact.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

Properties

The new list of properties. Overwrites the current property list.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 30 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2500.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 4096.

Value Pattern: `.*`

Required: No

PropertiesToRemove

A list of properties to remove.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 2500.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "ArtifactArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ArtifactArn

The Amazon Resource Name (ARN) of the artifact.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:artifact/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateCluster

Service: Amazon SageMaker Service

Updates a SageMaker HyperPod cluster.

Request Syntax

```
{
   "AutoScaling": {
      "AutoScalerType": "string",
      "Mode": "string"
   },
   "ClusterName": "string",
   "ClusterRole": "string",
   "InstanceGroups": [
      {
         "CapacityRequirements": {
            "OnDemand": {
            },
            "Spot": {
            }
         },
         "ExecutionRole": "string",
         "ImageId": "string",
         "InstanceCount": number,
         "InstanceGroupName": "string",
         "InstanceRequirements": {
            "InstanceTypes": [ "string" ]
         },
         "InstanceStorageConfigs": [
            { ... }
         ],
         "InstanceType": "string",
         "KubernetesConfig": {
            "Labels": {
               "string" : "string"
            },
            "Taints": [
               {
                  "Effect": "string",
                  "Key": "string",
                  "Value": "string"
               }
```

```
            ]
         },
         "LifeCycleConfig": {
            "OnCreate": "string",
            "OnInitComplete": "string",
            "SourceS3Uri": "string"
         },
         "MinInstanceCount": number,
         "NetworkInterface": {
            "InterfaceType": "string"
         },
         "OnStartDeepHealthChecks": [ "string" ],
         "OverrideVpcConfig": {
            "SecurityGroupIds": [ "string" ],
            "Subnets": [ "string" ]
         },
         "ScheduledUpdateConfig": {
            "DeploymentConfig": {
               "AutoRollbackConfiguration": [
                  {
                     "AlarmName": "string"
                  }
               ],
               "RollingUpdatePolicy": {
                  "MaximumBatchSize": {
                     "Type": "string",
                     "Value": number
                  },
                  "RollbackMaximumBatchSize": {
                     "Type": "string",
                     "Value": number
                  }
               },
               "WaitIntervalInSeconds": number
            },
            "ScheduleExpression": "string"
         },
         "SlurmConfig": {
            "NodeType": "string",
            "PartitionNames": [ "string" ]
         },
         "ThreadsPerCore": number,
         "TrainingPlanArn": "string"
      }
```

```
    ],
    "InstanceGroupsToDelete": [ "string" ],
    "NodeProvisioningMode": "string",
    "NodeRecovery": "string",
    "Orchestrator": {
       "Eks": {
          "ClusterArn": "string"
       },
       "Slurm": {
          "SlurmConfigStrategy": "string"
       }
    },
    "RestrictedInstanceGroups": [
       {
          "EnvironmentConfig": {
             "FSxLustreConfig": {
                "PerUnitStorageThroughput": number,
                "SizeInGiB": number
             }
          },
          "ExecutionRole": "string",
          "InstanceCount": number,
          "InstanceGroupName": "string",
          "InstanceStorageConfigs": [
             { ... }
          ],
          "InstanceType": "string",
          "OnStartDeepHealthChecks": [ "string" ],
          "OverrideVpcConfig": {
             "SecurityGroupIds": [ "string" ],
             "Subnets": [ "string" ]
          },
          "ScheduledUpdateConfig": {
             "DeploymentConfig": {
                "AutoRollbackConfiguration": [
                   {
                      "AlarmName": "string"
                   }
                ],
                "RollingUpdatePolicy": {
                   "MaximumBatchSize": {
                      "Type": "string",
                      "Value": number
                   },
```

```
                    "RollbackMaximumBatchSize": {
                       "Type": "string",
                       "Value": number
                    }
                 },
                 "WaitIntervalInSeconds": number
              },
              "ScheduleExpression": "string"
           },
           "ThreadsPerCore": number,
           "TrainingPlanArn": "string"
        }
    ],
    "RestrictedInstanceGroupsConfig": {
       "SharedEnvironmentConfig": {
          "FSxLustreConfig": {
             "PerUnitStorageThroughput": number,
             "SizeInGiB": number
          },
          "FSxLustreDeletionPolicy": "string"
       }
    },
    "TieredStorageConfig": {
       "InstanceMemoryAllocationPercentage": number,
       "Mode": "string"
    }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AutoScaling

Updates the autoscaling configuration for the cluster. Use to enable or disable automatic node scaling.

Type: ClusterAutoScalingConfig object

Required: No

ClusterName

Specify the name of the SageMaker HyperPod cluster you want to update.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

ClusterRole

The Amazon Resource Name (ARN) of the IAM role that HyperPod assumes for cluster autoscaling operations. Cannot be updated while autoscaling is enabled.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

InstanceGroups

Specify the instance groups to update.

Type: Array of ClusterInstanceGroupSpecification objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

InstanceGroupsToDelete

Specify the names of the instance groups to delete. Use a single `,` as the separator between multiple names.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

NodeProvisioningMode

Determines how instance provisioning is handled during cluster operations. In `Continuous` mode, the cluster provisions available instances incrementally and retries until the target count is reached. The cluster becomes operational once cluster-level resources are ready. Use `CurrentCount` and `TargetCount` in `DescribeCluster` to track provisioning progress.

Type: String

Valid Values: `Continuous`

Required: No

NodeRecovery

The node recovery mode to be applied to the SageMaker HyperPod cluster.

Type: String

Valid Values: `Automatic | None`

Required: No

Orchestrator

The type of orchestrator used for the SageMaker HyperPod cluster.

Type: ClusterOrchestrator object

Required: No

RestrictedInstanceGroups

The specialized instance groups for training models like Amazon Nova to be created in the SageMaker HyperPod cluster.

Type: Array of ClusterRestrictedInstanceGroupSpecification objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

RestrictedInstanceGroupsConfig

The configuration for the restricted instance groups (RIG) in the SageMaker HyperPod cluster.

Type: ClusterRestrictedInstanceGroupsConfig object

Required: No

TieredStorageConfig

Updates the configuration for managed tier checkpointing on the HyperPod cluster. For example, you can enable or disable the feature and modify the percentage of cluster memory allocated for checkpoint storage.

Type: ClusterTieredStorageConfig object

Required: No

Response Syntax

```
{
   "ClusterArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ClusterArn

The Amazon Resource Name (ARN) of the updated SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateClusterSchedulerConfig

Service: Amazon SageMaker Service

Update the cluster policy configuration.

Request Syntax

```
{
   "ClusterSchedulerConfigId": "string",
   "Description": "string",
   "SchedulerConfig": {
      "FairShare": "string",
      "IdleResourceSharing": "string",
      "PriorityClasses": [
         {
            "Name": "string",
            "Weight": number
         }
      ]
   },
   "TargetVersion": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterSchedulerConfigId

ID of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[a-z0-9]{12}`

Required: Yes

Description

Description of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

SchedulerConfig

Cluster policy configuration.

Type: SchedulerConfig object

Required: No

TargetVersion

Target version.

Type: Integer

Required: Yes

Response Syntax

```
{
   "ClusterSchedulerConfigArn": "string",
   "ClusterSchedulerConfigVersion": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ClusterSchedulerConfigArn

ARN of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:cluster-scheduler-config/[a-z0-9]{12}`

ClusterSchedulerConfigVersion

Version of the cluster policy.

Type: Integer

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateClusterSoftware

Service: Amazon SageMaker Service

Updates the platform software of a SageMaker HyperPod cluster for security patching. To learn how to use this API, see Update the SageMaker HyperPod platform software of a cluster.

> **Important**
>
> The `UpgradeClusterSoftware` API call may impact your SageMaker HyperPod cluster uptime and availability. Plan accordingly to mitigate potential disruptions to your workloads.

Request Syntax

```
{
   "ClusterName": "string",
   "DeploymentConfig": {
      "AutoRollbackConfiguration": [
         {
            "AlarmName": "string"
         }
      ],
      "RollingUpdatePolicy": {
         "MaximumBatchSize": {
            "Type": "string",
            "Value": number
         },
         "RollbackMaximumBatchSize": {
            "Type": "string",
            "Value": number
         }
      },
      "WaitIntervalInSeconds": number
   },
   "ImageId": "string",
   "InstanceGroups": [
      {
         "InstanceGroupName": "string"
      }
   ]
```

```
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ClusterName

Specify the name or the Amazon Resource Name (ARN) of the SageMaker HyperPod cluster you want to update for security patching.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

DeploymentConfig

The configuration to use when updating the AMI versions.

Type: DeploymentConfiguration object

Required: No

ImageId

When configuring your HyperPod cluster, you can specify an image ID using one of the following options:

- `HyperPodPublicAmiId`: Use a HyperPod public AMI
- `CustomAmiId`: Use your custom AMI
- `default`: Use the default latest system image

If you choose to use a custom AMI (`CustomAmiId`), ensure it meets the following requirements:

- Encryption: The custom AMI must be unencrypted.
- Ownership: The custom AMI must be owned by the same AWS account that is creating the HyperPod cluster.

- Volume support: Only the primary AMI snapshot volume is supported; additional AMI volumes are not supported.

When updating the instance group's AMI through the `UpdateClusterSoftware` operation, if an instance group uses a custom AMI, you must provide an `ImageId` or use the default as input. Note that if you don't specify an instance group in your `UpdateClusterSoftware` request, then all of the instance groups are patched with the specified image.

Type: String

Length Constraints: Minimum length of 7. Maximum length of 21.

Pattern: `ami-[0-9a-fA-F]{8,17}|default`

Required: No

InstanceGroups

The array of instance groups for which to update AMI versions.

Type: Array of UpdateClusterSoftwareInstanceGroupSpecification objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

Response Syntax

```
{
   "ClusterArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ClusterArn

The Amazon Resource Name (ARN) of the SageMaker HyperPod cluster being updated for security patching.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python

- AWS SDK for Ruby V3

UpdateCodeRepository

Service: Amazon SageMaker Service

Updates the specified Git repository with the specified values.

Request Syntax

```
{
   "CodeRepositoryName": "string",
   "GitConfig": {
      "SecretArn": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

CodeRepositoryName

The name of the Git repository to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

GitConfig

The configuration of the git repository, including the URL and the Amazon Resource Name (ARN) of the AWS Secrets Manager secret that contains the credentials used to access the repository. The secret must have a staging label of `AWSCURRENT` and must be in the following format:

`{"username": UserName, "password": Password}`

Type: GitConfigForUpdate object

Required: No

Response Syntax

```
{
   "CodeRepositoryArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CodeRepositoryArn

The ARN of the Git repository.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:code-repository/[\S]{1,2048}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateComputeQuota

Service: Amazon SageMaker Service

Update the compute allocation definition.

Request Syntax

```
{
   "ActivationState": "string",
   "ComputeQuotaConfig": {
      "ComputeQuotaResources": [
         {
            "AcceleratorPartition": {
               "Count": number,
               "Type": "string"
            },
            "Accelerators": number,
            "Count": number,
            "InstanceType": "string",
            "MemoryInGiB": number,
            "VCpu": number
         }
      ],
      "PreemptTeamTasks": "string",
      "ResourceSharingConfig": {
         "AbsoluteBorrowLimits": [
            {
               "AcceleratorPartition": {
                  "Count": number,
                  "Type": "string"
               },
               "Accelerators": number,
               "Count": number,
               "InstanceType": "string",
               "MemoryInGiB": number,
               "VCpu": number
            }
         ],
         "BorrowLimit": number,
         "Strategy": "string"
      }
   },
   "ComputeQuotaId": "string",
```

```
   "ComputeQuotaTarget": {
      "FairShareWeight": number,
      "TeamName": "string"
   },
   "Description": "string",
   "TargetVersion": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ActivationState

The state of the compute allocation being described. Use to enable or disable compute allocation.

Default is `Enabled`.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

ComputeQuotaConfig

Configuration of the compute allocation definition. This includes the resource sharing option, and the setting to preempt low priority tasks.

Type: ComputeQuotaConfig object

Required: No

ComputeQuotaId

ID of the compute allocation definition.

Type: String

Pattern: `[a-z0-9]{12}`

Required: Yes

ComputeQuotaTarget

The target entity to allocate compute resources to.

Type: ComputeQuotaTarget object

Required: No

Description

Description of the compute allocation definition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

TargetVersion

Target version.

Type: Integer

Required: Yes

Response Syntax

```
{
   "ComputeQuotaArn": "string",
   "ComputeQuotaVersion": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ComputeQuotaArn

ARN of the compute allocation definition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:compute-quota/[a-z0-9]{12}`

ComputeQuotaVersion

Version of the compute allocation definition.

Type: Integer

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateContext

Service: Amazon SageMaker Service

Updates a context.

Request Syntax

```
{
   "ContextName": "string",
   "Description": "string",
   "Properties": {
      "string" : "string"
   },
   "PropertiesToRemove": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ContextName

The name of the context to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,119}`

Required: Yes

Description

The new description for the context.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

Properties

The new list of properties. Overwrites the current property list.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 30 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2500.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Required: No

PropertiesToRemove

A list of properties to remove.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 2500.

Pattern: `.*`

Required: No

Response Syntax

```
{
   "ContextArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ContextArn

The Amazon Resource Name (ARN) of the context.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:context/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateDeviceFleet

Service: Amazon SageMaker Service

Updates a fleet of devices.

Request Syntax

```
{
   "Description": "string",
   "DeviceFleetName": "string",
   "EnableIotRoleAlias": boolean,
   "OutputConfig": {
      "KmsKeyId": "string",
      "PresetDeploymentConfig": "string",
      "PresetDeploymentType": "string",
      "S3OutputLocation": "string"
   },
   "RoleArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Description

Description of the fleet.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 800.

Pattern: `[-a-zA-Z0-9_.,;:! ]*`

Required: No

DeviceFleetName

The name of the fleet.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EnableIotRoleAlias

Whether to create an AWS IoT Role Alias during device fleet creation. The name of the role alias generated will match this pattern: "SageMakerEdge-{DeviceFleetName}".

For example, if your device fleet is called "demo-fleet", the name of the role alias will be "SageMakerEdge-demo-fleet".

Type: Boolean

Required: No

OutputConfig

Output configuration for storing sample data collected by the fleet.

Type: EdgeOutputConfig object

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of the device.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateDevices

Service: Amazon SageMaker Service

Updates one or more devices in a fleet.

Request Syntax

```
{
   "DeviceFleetName": "string",
   "Devices": [
      {
         "Description": "string",
         "DeviceName": "string",
         "IotThingName": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeviceFleetName

The name of the fleet the devices belong to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Devices

List of devices to register with Edge Manager agent.

Type: Array of Device objects

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateDomain

Service: Amazon SageMaker Service

Updates the default settings for new user profiles in the domain.

Request Syntax

```
{
   "AppNetworkAccessType": "string",
   "AppSecurityGroupManagement": "string",
   "DefaultSpaceSettings": {
      "CustomFileSystemConfigs": [
         { ... }
      ],
      "CustomPosixUserConfig": {
         "Gid": number,
         "Uid": number
      },
      "ExecutionRole": "string",
      "JupyterLabAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number,
               "LifecycleManagement": "string",
               "MaxIdleTimeoutInMinutes": number,
               "MinIdleTimeoutInMinutes": number
            }
         },
         "BuiltInLifecycleConfigArn": "string",
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
```

```
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "EmrSettings": {
            "AssumableRoleArns": [ "string" ],
            "ExecutionRoleArns": [ "string" ]
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "JupyterServerAppSettings": {
         "CodeRepositories": [
            {
               "RepositoryUrl": "string"
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "KernelGatewayAppSettings": {
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
```

```
            "LifecycleConfigArns": [ "string" ]
         },
         "SecurityGroups": [ "string" ],
         "SpaceStorageSettings": {
            "DefaultEbsStorageSettings": {
               "DefaultEbsVolumeSizeInGb": number,
               "MaximumEbsVolumeSizeInGb": number
            }
         }
      },
      "DefaultUserSettings": {
         "AutoMountHomeEFS": "string",
         "CanvasAppSettings": {
            "DirectDeploySettings": {
               "Status": "string"
            },
            "EmrServerlessSettings": {
               "ExecutionRoleArn": "string",
               "Status": "string"
            },
            "GenerativeAiSettings": {
               "AmazonBedrockRoleArn": "string"
            },
            "IdentityProviderOAuthSettings": [
               {
                  "DataSourceName": "string",
                  "SecretArn": "string",
                  "Status": "string"
               }
            ],
            "KendraSettings": {
               "Status": "string"
            },
            "ModelRegisterSettings": {
               "CrossAccountModelRegisterRoleArn": "string",
               "Status": "string"
            },
            "TimeSeriesForecastingSettings": {
               "AmazonForecastRoleArn": "string",
               "Status": "string"
            },
            "WorkspaceSettings": {
               "S3ArtifactPath": "string",
               "S3KmsKeyId": "string"
```

```
      }
   },
   "CodeEditorAppSettings": {
      "AppLifecycleManagement": {
         "IdleSettings": {
            "IdleTimeoutInMinutes": number,
            "LifecycleManagement": "string",
            "MaxIdleTimeoutInMinutes": number,
            "MinIdleTimeoutInMinutes": number
         }
      },
      "BuiltInLifecycleConfigArn": "string",
      "CustomImages": [
         {
            "AppImageConfigName": "string",
            "ImageName": "string",
            "ImageVersionNumber": number
         }
      ],
      "DefaultResourceSpec": {
         "InstanceType": "string",
         "LifecycleConfigArn": "string",
         "SageMakerImageArn": "string",
         "SageMakerImageVersionAlias": "string",
         "SageMakerImageVersionArn": "string",
         "TrainingPlanArn": "string"
      },
      "LifecycleConfigArns": [ "string" ]
   },
   "CustomFileSystemConfigs": [
      { ... }
   ],
   "CustomPosixUserConfig": {
      "Gid": number,
      "Uid": number
   },
   "DefaultLandingUri": "string",
   "ExecutionRole": "string",
   "JupyterLabAppSettings": {
      "AppLifecycleManagement": {
         "IdleSettings": {
            "IdleTimeoutInMinutes": number,
            "LifecycleManagement": "string",
            "MaxIdleTimeoutInMinutes": number,
```

```
            "MinIdleTimeoutInMinutes": number
         }
      },
      "BuiltInLifecycleConfigArn": "string",
      "CodeRepositories": [
         {
            "RepositoryUrl": "string"
         }
      ],
      "CustomImages": [
         {
            "AppImageConfigName": "string",
            "ImageName": "string",
            "ImageVersionNumber": number
         }
      ],
      "DefaultResourceSpec": {
         "InstanceType": "string",
         "LifecycleConfigArn": "string",
         "SageMakerImageArn": "string",
         "SageMakerImageVersionAlias": "string",
         "SageMakerImageVersionArn": "string",
         "TrainingPlanArn": "string"
      },
      "EmrSettings": {
         "AssumableRoleArns": [ "string" ],
         "ExecutionRoleArns": [ "string" ]
      },
      "LifecycleConfigArns": [ "string" ]
   },
   "JupyterServerAppSettings": {
      "CodeRepositories": [
         {
            "RepositoryUrl": "string"
         }
      ],
      "DefaultResourceSpec": {
         "InstanceType": "string",
         "LifecycleConfigArn": "string",
         "SageMakerImageArn": "string",
         "SageMakerImageVersionAlias": "string",
         "SageMakerImageVersionArn": "string",
         "TrainingPlanArn": "string"
      },
```

```
            "LifecycleConfigArns": [ "string" ]
         },
         "KernelGatewayAppSettings": {
            "CustomImages": [
               {
                  "AppImageConfigName": "string",
                  "ImageName": "string",
                  "ImageVersionNumber": number
               }
            ],
            "DefaultResourceSpec": {
               "InstanceType": "string",
               "LifecycleConfigArn": "string",
               "SageMakerImageArn": "string",
               "SageMakerImageVersionAlias": "string",
               "SageMakerImageVersionArn": "string",
               "TrainingPlanArn": "string"
            },
            "LifecycleConfigArns": [ "string" ]
         },
         "RSessionAppSettings": {
            "CustomImages": [
               {
                  "AppImageConfigName": "string",
                  "ImageName": "string",
                  "ImageVersionNumber": number
               }
            ],
            "DefaultResourceSpec": {
               "InstanceType": "string",
               "LifecycleConfigArn": "string",
               "SageMakerImageArn": "string",
               "SageMakerImageVersionAlias": "string",
               "SageMakerImageVersionArn": "string",
               "TrainingPlanArn": "string"
            }
         },
         "RStudioServerProAppSettings": {
            "AccessStatus": "string",
            "UserGroup": "string"
         },
         "SecurityGroups": [ "string" ],
         "SharingSettings": {
            "NotebookOutputOption": "string",
```

```
            "S3KmsKeyId": "string",
            "S3OutputPath": "string"
         },
         "SpaceStorageSettings": {
            "DefaultEbsStorageSettings": {
               "DefaultEbsVolumeSizeInGb": number,
               "MaximumEbsVolumeSizeInGb": number
            }
         },
         "StudioWebPortal": "string",
         "StudioWebPortalSettings": {
            "ExecutionRoleSessionNameMode": "string",
            "HiddenAppTypes": [ "string" ],
            "HiddenInstanceTypes": [ "string" ],
            "HiddenMlTools": [ "string" ],
            "HiddenSageMakerImageVersionAliases": [
               {
                  "SageMakerImageName": "string",
                  "VersionAliases": [ "string" ]
               }
            ]
         },
         "TensorBoardAppSettings": {
            "DefaultResourceSpec": {
               "InstanceType": "string",
               "LifecycleConfigArn": "string",
               "SageMakerImageArn": "string",
               "SageMakerImageVersionAlias": "string",
               "SageMakerImageVersionArn": "string",
               "TrainingPlanArn": "string"
            }
         }
      },
      "DomainId": "string",
      "DomainSettingsForUpdate": {
         "AmazonQSettings": {
            "QProfileArn": "string",
            "Status": "string"
         },
         "DockerSettings": {
            "EnableDockerAccess": "string",
            "RootlessDocker": "string",
            "VpcOnlyTrustedAccounts": [ "string" ]
         },
```

```
      "ExecutionRoleIdentityConfig": "string",
      "IpAddressType": "string",
      "RStudioServerProDomainSettingsForUpdate": {
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "DomainExecutionRoleArn": "string",
         "RStudioConnectUrl": "string",
         "RStudioPackageManagerUrl": "string"
      },
      "SecurityGroupIds": [ "string" ],
      "TrustedIdentityPropagationSettings": {
         "Status": "string"
      },
      "UnifiedStudioSettings": {
         "DomainAccountId": "string",
         "DomainId": "string",
         "DomainRegion": "string",
         "EnvironmentId": "string",
         "ProjectId": "string",
         "ProjectS3Path": "string",
         "SingleSignOnApplicationArn": "string",
         "StudioWebPortalAccess": "string"
      }
   },
   "HomeEfsFileSystemCreation": "string",
   "SubnetIds": [ "string" ],
   "TagPropagation": "string",
   "VpcId": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AppNetworkAccessType

Specifies the VPC used for non-EFS traffic.

- `PublicInternetOnly` - Non-EFS traffic is through a VPC managed by Amazon SageMaker AI, which allows direct internet access.
- `VpcOnly` - All Studio traffic is through the specified VPC and subnets.

This configuration can only be modified if there are no apps in the `InService`, `Pending`, or `Deleting` state. The configuration cannot be updated if `DomainSettings.RStudioServerProDomainSettings.DomainExecutionRoleArn` is already set or `DomainSettings.RStudioServerProDomainSettings.DomainExecutionRoleArn` is provided as part of the same request.

Type: String

Valid Values: `PublicInternetOnly | VpcOnly`

Required: No

AppSecurityGroupManagement

The entity that creates and manages the required security groups for inter-app communication in `VPCOnly` mode. Required when `CreateDomain.AppNetworkAccessType` is `VPCOnly` and `DomainSettings.RStudioServerProDomainSettings.DomainExecutionRoleArn` is provided. If setting up the domain for use with RStudio, this value must be set to `Service`.

Type: String

Valid Values: `Service | Customer`

Required: No

DefaultSpaceSettings

The default settings for shared spaces that users create in the domain.

Type: DefaultSpaceSettings object

Required: No

DefaultUserSettings

A collection of settings.

Type: UserSettings object

Required: No

DomainId

The ID of the domain to be updated.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

DomainSettingsForUpdate

A collection of `DomainSettings` configuration values to update.

Type: DomainSettingsForUpdate object

Required: No

HomeEfsFileSystemCreation

Indicates whether to create a home EFS file system for the domain. You can change from `Disabled` to `Enabled` to provision EFS on demand, but you cannot change from `Enabled` to `Disabled`.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

SubnetIds

The VPC subnets that Studio uses for communication.

If removing subnets, ensure there are no apps in the `InService`, `Pending`, or `Deleting` state.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: No

TagPropagation

Indicates whether custom tag propagation is supported for the domain. Defaults to `DISABLED`.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

VpcId

The identifier for the VPC used by the domain for network communication. Use this field only when adding VPC configuration to a SageMaker AI domain used in Amazon SageMaker Unified Studio that was created without VPC settings. SageMaker AI doesn't automatically apply VPC updates to existing applications. Stop and restart your applications to apply the changes.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: No

Response Syntax

```
{
   "DomainArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

DomainArn

The Amazon Resource Name (ARN) of the domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:domain/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateEndpoint

Service: Amazon SageMaker Service

Deploys the `EndpointConfig` specified in the request to a new fleet of instances. SageMaker shifts endpoint traffic to the new instances with the updated endpoint configuration and then deletes the old instances using the previous `EndpointConfig` (there is no availability loss). For more information about how to control the update and traffic shifting process, see Update models in production.

When SageMaker receives the request, it sets the endpoint status to `Updating`. After updating the endpoint, it sets the status to `InService`. To check the status of an endpoint, use the DescribeEndpoint API.

> **Note**
>
> You must not delete an `EndpointConfig` in use by an endpoint that is live or while the `UpdateEndpoint` or `CreateEndpoint` operations are being performed on the endpoint. To update an endpoint, you must create a new `EndpointConfig`.
> If you delete the `EndpointConfig` of an endpoint that is active or being created or updated you may lose visibility into the instance type the endpoint is using. The endpoint must be deleted in order to stop incurring charges.

Request Syntax

```
{
   "DeploymentConfig": {
      "AutoRollbackConfiguration": {
         "Alarms": [
            {
               "AlarmName": "string"
            }
         ]
      },
      "BlueGreenUpdatePolicy": {
         "MaximumExecutionTimeoutInSeconds": number,
         "TerminationWaitInSeconds": number,
         "TrafficRoutingConfiguration": {
            "CanarySize": {
               "Type": "string",
```

```
                "Value": number
            },
            "LinearStepSize": {
                "Type": "string",
                "Value": number
            },
            "Type": "string",
            "WaitIntervalInSeconds": number
        }
    },
    "RollingUpdatePolicy": {
        "MaximumBatchSize": {
            "Type": "string",
            "Value": number
        },
        "MaximumExecutionTimeoutInSeconds": number,
        "RollbackMaximumBatchSize": {
            "Type": "string",
            "Value": number
        },
        "WaitIntervalInSeconds": number
    }
},
"EndpointConfigName": "string",
"EndpointName": "string",
"ExcludeRetainedVariantProperties": [
    {
        "VariantPropertyType": "string"
    }
],
"RetainAllVariantProperties": boolean,
"RetainDeploymentConfig": boolean
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeploymentConfig

The deployment configuration for an endpoint, which contains the desired deployment strategy and rollback configurations.

Type: DeploymentConfig object

Required: No

EndpointConfigName

The name of the new endpoint configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EndpointName

The name of the endpoint whose configuration you want to update.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ExcludeRetainedVariantProperties

When you are updating endpoint resources with `RetainAllVariantProperties`, whose value is set to `true`, `ExcludeRetainedVariantProperties` specifies the list of type VariantProperty to override with the values provided by `EndpointConfig`. If you don't specify a value for `ExcludeRetainedVariantProperties`, no variant properties are overridden.

Type: Array of VariantProperty objects

Array Members: Minimum number of 0 items. Maximum number of 3 items.

Required: No

RetainAllVariantProperties

When updating endpoint resources, enables or disables the retention of variant properties, such as the instance count or the variant weight. To retain the variant properties of an endpoint when updating it, set `RetainAllVariantProperties` to `true`. To use the

variant properties specified in a new `EndpointConfig` call when updating an endpoint, set `RetainAllVariantProperties` to `false`. The default is `false`.

Type: Boolean

Required: No

RetainDeploymentConfig

Specifies whether to reuse the last deployment configuration. The default value is false (the configuration is not reused).

Type: Boolean

Required: No

Response Syntax

```
{
   "EndpointArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

EndpointArn

The Amazon Resource Name (ARN) of the endpoint.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateEndpointWeightsAndCapacities

Service: Amazon SageMaker Service

Updates variant weight of one or more variants associated with an existing endpoint, or capacity of one variant associated with an existing endpoint. When it receives the request, SageMaker sets the endpoint status to `Updating`. After updating the endpoint, it sets the status to `InService`. To check the status of an endpoint, use the DescribeEndpoint API.

Request Syntax

```
{
   "DesiredWeightsAndCapacities": [
      {
         "DesiredInstanceCount": number,
         "DesiredWeight": number,
         "ServerlessUpdateConfig": {
            "MaxConcurrency": number,
            "ProvisionedConcurrency": number
         },
         "VariantName": "string"
      }
   ],
   "EndpointName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DesiredWeightsAndCapacities

An object that provides new capacity and weight values for a variant.

Type: Array of DesiredWeightAndCapacity objects

Array Members: Minimum number of 1 item.

Required: Yes

EndpointName

The name of an existing SageMaker endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "EndpointArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

EndpointArn

The Amazon Resource Name (ARN) of the updated endpoint.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateExperiment

Service: Amazon SageMaker Service

Adds, updates, or removes the description of an experiment. Updates the display name of an experiment.

Request Syntax

```
{
   "Description": "string",
   "DisplayName": "string",
   "ExperimentName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Description

The description of the experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

DisplayName

The name of the experiment as displayed. The name doesn't need to be unique. If `DisplayName` isn't specified, `ExperimentName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

ExperimentName

The name of the experiment to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "ExperimentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ExperimentArn

The Amazon Resource Name (ARN) of the experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateFeatureGroup

Service: Amazon SageMaker Service

Updates the feature group by either adding features or updating the online store configuration. Use one of the following request parameters at a time while using the `UpdateFeatureGroup` API.

You can add features for your feature group using the `FeatureAdditions` request parameter. Features cannot be removed from a feature group.

You can update the online store configuration by using the `OnlineStoreConfig` request parameter. If a `TtlDuration` is specified, the default `TtlDuration` applies for all records added to the feature group *after the feature group is updated*. If a record level `TtlDuration` exists from using the `PutRecord` API, the record level `TtlDuration` applies to that record instead of the default `TtlDuration`. To remove the default `TtlDuration` from an existing feature group, use the `UpdateFeatureGroup` API and set the `TtlDuration` `Unit` and `Value` to `null`.

Request Syntax

```
{
   "FeatureAdditions": [
      {
         "CollectionConfig": { ... },
         "CollectionType": "string",
         "FeatureName": "string",
         "FeatureType": "string"
      }
   ],
   "FeatureGroupName": "string",
   "OnlineStoreConfig": {
      "TtlDuration": {
         "Unit": "string",
         "Value": number
      }
   },
   "ThroughputConfig": {
      "ProvisionedReadCapacityUnits": number,
      "ProvisionedWriteCapacityUnits": number,
      "ThroughputMode": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

FeatureAdditions

Updates the feature group. Updating a feature group is an asynchronous operation. When you get an HTTP 200 response, you've made a valid request. It takes some time after you've made a valid request for Feature Store to update the feature group.

Type: Array of FeatureDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

FeatureGroupName

The name or Amazon Resource Name (ARN) of the feature group that you're updating.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group\/)?([a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63})`

Required: Yes

OnlineStoreConfig

Updates the feature group online store configuration.

Type: OnlineStoreConfigUpdate object

Required: No

ThroughputConfig

The new throughput configuration for the feature group. You can switch between on-demand and provisioned modes or update the read / write capacity of provisioned feature groups. You can switch a feature group to on-demand only once in a 24 hour period.

Type: ThroughputConfigUpdate object

Required: No

Response Syntax

```
{
   "FeatureGroupArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

FeatureGroupArn

The Amazon Resource Number (ARN) of the feature group that you're updating.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateFeatureMetadata

Service: Amazon SageMaker Service

Updates the description and parameters of the feature group.

Request Syntax

```
{
   "Description": "string",
   "FeatureGroupName": "string",
   "FeatureName": "string",
   "ParameterAdditions": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "ParameterRemovals": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Description

A description that you can write to better describe the feature.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

FeatureGroupName

The name or Amazon Resource Name (ARN) of the feature group containing the feature that you're updating.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group\/)?([a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63})`

Required: Yes

FeatureName

The name of the feature that you're updating.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

Required: Yes

ParameterAdditions

A list of key-value pairs that you can add to better describe the feature.

Type: Array of FeatureParameter objects

Array Members: Minimum number of 0 items. Maximum number of 25 items.

Required: No

ParameterRemovals

A list of parameter keys that you can specify to remove parameters that describe your feature.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 25 items.

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `([\p{L}\p{Z}\p{N}_.:/=+\-]*)`

Required: No

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateHub

Service: Amazon SageMaker Service

Update a hub.

Request Syntax

```
{
   "HubDescription": "string",
   "HubDisplayName": "string",
   "HubName": "string",
   "HubSearchKeywords": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HubDescription

A description of the updated hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Pattern: `.*`

Required: No

HubDisplayName

The display name of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

HubName

The name of the hub to update.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubSearchKeywords

The searchable keywords for the hub.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[^A-Z]*`

Required: No

Response Syntax

```
{
   "HubArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HubArn

The Amazon Resource Name (ARN) of the updated hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateHubContent

Service: Amazon SageMaker Service

Updates SageMaker hub content (either a `Model` or `Notebook` resource).

You can update the metadata that describes the resource. In addition to the required request fields, specify at least one of the following fields to update:

- `HubContentDescription`
- `HubContentDisplayName`
- `HubContentMarkdown`
- `HubContentSearchKeywords`
- `SupportStatus`

For more information about hubs, see Private curated hubs for foundation model access control in JumpStart.

> **Note**
>
> If you want to update a `ModelReference` resource in your hub, use the `UpdateHubContentResource` API instead.

Request Syntax

```
{
   "HubContentDescription": "string",
   "HubContentDisplayName": "string",
   "HubContentMarkdown": "string",
   "HubContentName": "string",
   "HubContentSearchKeywords": [ "string" ],
   "HubContentType": "string",
   "HubContentVersion": "string",
   "HubName": "string",
   "SupportStatus": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HubContentDescription

The description of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Pattern: `.*`

Required: No

HubContentDisplayName

The display name of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

HubContentMarkdown

A string that provides a description of the hub content. This string can include links, tables, and standard markdown formatting.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 170391.

Required: No

HubContentName

The name of the hub content resource that you want to update.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubContentSearchKeywords

The searchable keywords of the hub content.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

HubContentType

The content type of the resource that you want to update. Only specify a `Model` or `Notebook` resource for this API. To update a `ModelReference`, use the `UpdateHubContentReference` API instead.

Type: String

Valid Values: `Model | Notebook | ModelReference | DataSet | JsonDoc`

Required: Yes

HubContentVersion

The hub content version that you want to update. For example, if you have two versions of a resource in your hub, you can update the second version.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: Yes

HubName

The name of the SageMaker hub that contains the hub content you want to update. You can optionally use the hub ARN instead.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

SupportStatus

Indicates the current status of the hub content resource.

Type: String

Valid Values: `Supported | Deprecated | Restricted`

Required: No

Response Syntax

```
{
   "HubArn": "string",
   "HubContentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HubArn

The ARN of the private model hub that contains the updated hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

HubContentArn

The ARN of the hub content resource that was updated.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateHubContentReference

Service: Amazon SageMaker Service

Updates the contents of a SageMaker hub for a `ModelReference` resource. A `ModelReference` allows you to access public SageMaker JumpStart models from within your private hub.

When using this API, you can update the `MinVersion` field for additional flexibility in the model version. You shouldn't update any additional fields when using this API, because the metadata in your private hub should match the public JumpStart model's metadata.

> **Note**
>
> If you want to update a `Model` or `Notebook` resource in your hub, use the `UpdateHubContent` API instead.

For more information about adding model references to your hub, see Add models to a private hub.

Request Syntax

```
{
   "HubContentName": "string",
   "HubContentType": "string",
   "HubName": "string",
   "MinVersion": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

HubContentName

The name of the hub content resource that you want to update.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubContentType

The content type of the resource that you want to update. Only specify a `ModelReference` resource for this API. To update a `Model` or `Notebook` resource, use the `UpdateHubContent` API instead.

Type: String

Valid Values: `Model | Notebook | ModelReference | DataSet | JsonDoc`

Required: Yes

HubName

The name of the SageMaker hub that contains the hub content you want to update. You can optionally use the hub ARN instead.

Type: String

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub\/)?[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

MinVersion

The minimum hub content version of the referenced model that you want to use. The minimum version must be older than the latest available version of the referenced model. To support all versions of a model, set the value to `1.0.0`.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: No

Response Syntax

```
{
```

```
   "HubArn": "string",
   "HubContentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

HubArn

The ARN of the private model hub that contains the updated hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

HubContentArn

The ARN of the hub content resource that was updated.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateImage

Service: Amazon SageMaker Service

Updates the properties of a SageMaker AI image. To change the image's tags, use the AddTags and DeleteTags APIs.

Request Syntax

```
{
   "DeleteProperties": [ "string" ],
   "Description": "string",
   "DisplayName": "string",
   "ImageName": "string",
   "RoleArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeleteProperties

A list of properties to delete. Only the `Description` and `DisplayName` properties can be deleted.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2 items.

Length Constraints: Minimum length of 1. Maximum length of 11.

Pattern: `(^DisplayName$)|(^Description$)`

Required: No

Description

The new description for the image.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 512.

Pattern: `.*`

Required: No

DisplayName

The new display name for the image.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `\S(.*\S)?`

Required: No

ImageName

The name of the image to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

RoleArn

The new ARN for the IAM role that enables Amazon SageMaker AI to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

Response Syntax

```
{
```

```
   "ImageArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ImageArn

The ARN of the image.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image/[a-zA-Z0-9]([-.]?[a-zA-Z0-9])*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateImageVersion

Service: Amazon SageMaker Service

Updates the properties of a SageMaker AI image version.

Request Syntax

```
{
   "Alias": "string",
   "AliasesToAdd": [ "string" ],
   "AliasesToDelete": [ "string" ],
   "Horovod": boolean,
   "ImageName": "string",
   "JobType": "string",
   "MLFramework": "string",
   "Processor": "string",
   "ProgrammingLang": "string",
   "ReleaseNotes": "string",
   "VendorGuidance": "string",
   "Version": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Alias

The alias of the image version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(?!^[.-])^([a-zA-Z0-9-_.]+)`

Required: No

AliasesToAdd

A list of aliases to add.

Type: Array of strings

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(?!^[.-])^([a-zA-Z0-9-_.]+)`

Required: No

AliasesToDelete

A list of aliases to delete.

Type: Array of strings

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(?!^[.-])^([a-zA-Z0-9-_.]+)`

Required: No

Horovod

Indicates Horovod compatibility.

Type: Boolean

Required: No

ImageName

The name of the image.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

JobType

Indicates SageMaker AI job type compatibility.

- `TRAINING`: The image version is compatible with SageMaker AI training jobs.
- `INFERENCE`: The image version is compatible with SageMaker AI inference jobs.

- NOTEBOOK_KERNEL: The image version is compatible with SageMaker AI notebook kernels.

Type: String

Valid Values: TRAINING | INFERENCE | NOTEBOOK_KERNEL

Required: No

MLFramework

The machine learning framework vended in the image version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: [a-zA-Z]+ ?\d+\.\d+(\.\d+)?

Required: No

Processor

Indicates CPU or GPU compatibility.

- CPU: The image version is compatible with CPU.
- GPU: The image version is compatible with GPU.

Type: String

Valid Values: CPU | GPU

Required: No

ProgrammingLang

The supported programming language and its version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: [a-zA-Z]+ ?\d+\.\d+(\.\d+)?

Required: No

ReleaseNotes

The maintainer description of the image version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.*`

Required: No

VendorGuidance

The availability of the image version specified by the maintainer.

- `NOT_PROVIDED`: The maintainers did not provide a status for image version stability.
- `STABLE`: The image version is stable.
- `TO_BE_ARCHIVED`: The image version is set to be archived. Custom image versions that are set to be archived are automatically archived after three months.
- `ARCHIVED`: The image version is archived. Archived image versions are not searchable and are no longer actively supported.

Type: String

Valid Values: `NOT_PROVIDED | STABLE | TO_BE_ARCHIVED | ARCHIVED`

Required: No

Version

The version of the image.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

Response Syntax

```
{
```

```
   "ImageVersionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ImageVersionArn

The ARN of the image version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image-version/[a-z0-9]([-.]?[a-z0-9])*/[0-9]+|None)`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateInferenceComponent

Service: Amazon SageMaker Service

Updates an inference component.

Request Syntax

```
{
   "DeploymentConfig": {
      "AutoRollbackConfiguration": {
         "Alarms": [
            {
               "AlarmName": "string"
            }
         ]
      },
      "RollingUpdatePolicy": {
         "MaximumBatchSize": {
            "Type": "string",
            "Value": number
         },
         "MaximumExecutionTimeoutInSeconds": number,
         "RollbackMaximumBatchSize": {
            "Type": "string",
            "Value": number
         },
         "WaitIntervalInSeconds": number
      }
   },
   "InferenceComponentName": "string",
   "RuntimeConfig": {
      "CopyCount": number
   },
   "Specification": {
      "BaseInferenceComponentName": "string",
      "ComputeResourceRequirements": {
         "MaxMemoryRequiredInMb": number,
         "MinMemoryRequiredInMb": number,
         "NumberOfAcceleratorDevicesRequired": number,
         "NumberOfCpuCoresRequired": number
      },
      "Container": {
         "ArtifactUrl": "string",
```

```
            "ContainerMetricsConfig": {
               "MetricsEndpoints": [
                  {
                     "MetricPublishFrequencyInSeconds": number,
                     "MetricsEndpointPath": "string"
                  }
               ]
            },
            "Environment": {
               "string" : "string"
            },
            "Image": "string"
         },
         "DataCacheConfig": {
            "EnableCaching": boolean
         },
         "InstanceType": "string",
         "ModelName": "string",
         "SchedulingConfig": {
            "AvailabilityZoneBalance": {
               "EnforcementMode": "string",
               "MaxImbalance": number
            },
            "PlacementStrategy": "string"
         },
         "StartupParameters": {
            "ContainerStartupHealthCheckTimeoutInSeconds": number,
            "ModelDataDownloadTimeoutInSeconds": number
         }
      },
      "Specifications": [
         {
            "BaseInferenceComponentName": "string",
            "ComputeResourceRequirements": {
               "MaxMemoryRequiredInMb": number,
               "MinMemoryRequiredInMb": number,
               "NumberOfAcceleratorDevicesRequired": number,
               "NumberOfCpuCoresRequired": number
            },
            "Container": {
               "ArtifactUrl": "string",
               "ContainerMetricsConfig": {
                  "MetricsEndpoints": [
                     {
```

```
                    "MetricPublishFrequencyInSeconds": number,
                    "MetricsEndpointPath": "string"
                 }
              ]
           },
           "Environment": {
              "string" : "string"
           },
           "Image": "string"
        },
        "DataCacheConfig": {
           "EnableCaching": boolean
        },
        "InstanceType": "string",
        "ModelName": "string",
        "SchedulingConfig": {
           "AvailabilityZoneBalance": {
              "EnforcementMode": "string",
              "MaxImbalance": number
           },
           "PlacementStrategy": "string"
        },
        "StartupParameters": {
           "ContainerStartupHealthCheckTimeoutInSeconds": number,
           "ModelDataDownloadTimeoutInSeconds": number
        }
     }
  ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DeploymentConfig

The deployment configuration for the inference component. The configuration contains the desired deployment strategy and rollback settings.

Type: InferenceComponentDeploymentConfig object

Required: No

InferenceComponentName

The name of the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

RuntimeConfig

Runtime settings for a model that is deployed with an inference component.

Type: InferenceComponentRuntimeConfig object

Required: No

Specification

Details about the resources to deploy with this inference component, including the model, container, and compute resources.

Type: InferenceComponentSpecification object

Required: No

Specifications

A list of specification objects for the inference component, one per instance type. Use this parameter when you want to specify different model or resource configurations for the inference component on each instance type. You can use either this parameter or the singular `Specification` parameter, but not both.

Type: Array of InferenceComponentSpecification objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

Response Syntax

```
{
```

```
   "InferenceComponentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

InferenceComponentArn

The Amazon Resource Name (ARN) of the inference component.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateInferenceComponentRuntimeConfig

Service: Amazon SageMaker Service

Runtime settings for a model that is deployed with an inference component.

Request Syntax

```
{
   "DesiredRuntimeConfig": {
      "CopyCount": number
   },
   "InferenceComponentName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DesiredRuntimeConfig

Runtime settings for a model that is deployed with an inference component.

Type: InferenceComponentRuntimeConfig object

Required: Yes

InferenceComponentName

The name of the inference component to update.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

Response Syntax

```
{
```

```
   "InferenceComponentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

InferenceComponentArn

The Amazon Resource Name (ARN) of the inference component.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateInferenceExperiment

Service: Amazon SageMaker Service

Updates an inference experiment that you created. The status of the inference experiment has to be either `Created`, `Running`. For more information on the status of an inference experiment, see DescribeInferenceExperiment.

Request Syntax

```
{
   "DataStorageConfig": {
      "ContentType": {
         "CsvContentTypes": [ "string" ],
         "JsonContentTypes": [ "string" ]
      },
      "Destination": "string",
      "KmsKey": "string"
   },
   "Description": "string",
   "ModelVariants": [
      {
         "InfrastructureConfig": {
            "InfrastructureType": "string",
            "RealTimeInferenceConfig": {
               "InstanceCount": number,
               "InstanceType": "string"
            }
         },
         "ModelName": "string",
         "VariantName": "string"
      }
   ],
   "Name": "string",
   "Schedule": {
      "EndTime": number,
      "StartTime": number
   },
   "ShadowModeConfig": {
      "ShadowModelVariants": [
         {
            "SamplingPercentage": number,
            "ShadowModelVariantName": "string"
         }
```

```
        ],
        "SourceModelVariantName": "string"
    }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DataStorageConfig

The Amazon S3 location and configuration for storing inference request and response data.

Type: InferenceExperimentDataStorageConfig object

Required: No

Description

The description of the inference experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

ModelVariants

An array of `ModelVariantConfig` objects. There is one for each variant, whose infrastructure configuration you want to update.

Type: Array of ModelVariantConfig objects

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Required: No

Name

The name of the inference experiment to be updated.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Schedule

The duration for which the inference experiment will run. If the status of the inference experiment is `Created`, then you can update both the start and end dates. If the status of the inference experiment is `Running`, then you can update only the end date.

Type: InferenceExperimentSchedule object

Required: No

ShadowModeConfig

The configuration of `ShadowMode` inference experiment type. Use this field to specify a production variant which takes all the inference requests, and a shadow variant to which Amazon SageMaker replicates a percentage of the inference requests. For the shadow variant also specify the percentage of requests that Amazon SageMaker replicates.

Type: ShadowModeConfig object

Required: No

Response Syntax

```
{
   "InferenceExperimentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

InferenceExperimentArn

The ARN of the updated inference experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:inference-experiment/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python

- AWS SDK for Ruby V3

UpdateMlflowApp

Service: Amazon SageMaker Service

Updates an MLflow App.

Request Syntax

```
{
   "AccountDefaultStatus": "string",
   "Arn": "string",
   "ArtifactStoreUri": "string",
   "DefaultDomainIdList": [ "string" ],
   "ModelRegistrationMode": "string",
   "Name": "string",
   "WeeklyMaintenanceWindowStart": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AccountDefaultStatus

Indicates whether this this MLflow App is the default for the account.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

Arn

The ARN of the MLflow App to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-app/.*`

Required: Yes

ArtifactStoreUri

The new S3 URI for the general purpose bucket to use as the artifact store for the MLflow App.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

DefaultDomainIdList

List of SageMaker Domain IDs for which this MLflow App is the default.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: No

ModelRegistrationMode

Whether to enable or disable automatic registration of new MLflow models to the SageMaker Model Registry. To enable automatic model registration, set this value to `AutoModelRegistrationEnabled`. To disable automatic model registration, set this value to `AutoModelRegistrationDisabled`. If not specified, `AutomaticModelRegistration` defaults to `AutoModelRegistrationEnabled`

Type: String

Valid Values: `AutoModelRegistrationEnabled | AutoModelRegistrationDisabled`

Required: No

Name

The name of the MLflow App to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: No

WeeklyMaintenanceWindowStart

The new weekly maintenance window start day and time to update. The maintenance window day and time should be in Coordinated Universal Time (UTC) 24-hour standard time. For example: TUE:03:30.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 9.

Pattern: `(Mon|Tue|Wed|Thu|Fri|Sat|Sun):([01]\d|2[0-3]):([0-5]\d)`

Required: No

Response Syntax

```
{
   "Arn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The ARN of the updated MLflow App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-app/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateMlflowTrackingServer

Service: Amazon SageMaker Service

Updates properties of an existing MLflow Tracking Server.

Request Syntax

```
{
   "ArtifactStoreUri": "string",
   "AutomaticModelRegistration": boolean,
   "S3BucketOwnerAccountId": "string",
   "S3BucketOwnerVerification": boolean,
   "TrackingServerName": "string",
   "TrackingServerSize": "string",
   "WeeklyMaintenanceWindowStart": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ArtifactStoreUri

The new S3 URI for the general purpose bucket to use as the artifact store for the MLflow Tracking Server.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

AutomaticModelRegistration

Whether to enable or disable automatic registration of new MLflow models to the SageMaker Model Registry. To enable automatic model registration, set this value to `True`. To disable automatic model registration, set this value to `False`. If not specified, `AutomaticModelRegistration` defaults to `False`

Type: Boolean

Required: No

S3BucketOwnerAccountId

The new expected AWS account ID that owns the Amazon S3 bucket for artifact storage.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `\d+`

Required: No

S3BucketOwnerVerification

Whether to enable or disable Amazon S3 Bucket Owenrship Verifaction whenever the MLflow Tracking Server interacts with Amazon Amazon S3.

Type: Boolean

Required: No

TrackingServerName

The name of the MLflow Tracking Server to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: Yes

TrackingServerSize

The new size for the MLflow Tracking Server.

Type: String

Valid Values: `Small | Medium | Large`

Required: No

WeeklyMaintenanceWindowStart

The new weekly maintenance window start day and time to update. The maintenance window day and time should be in Coordinated Universal Time (UTC) 24-hour standard time. For example: TUE:03:30.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 9.

Pattern: `(Mon|Tue|Wed|Thu|Fri|Sat|Sun):([01]\d|2[0-3]):([0-5]\d)`

Required: No

Response Syntax

```
{
   "TrackingServerArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrackingServerArn

The ARN of the updated MLflow Tracking Server.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-tracking-server/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateModelCard

Service: Amazon SageMaker Service

Update an Amazon SageMaker Model Card.

> **Important**
>
> You cannot update both model card content and model card status in a single call.

Request Syntax

```
{
   "Content": "string",
   "ModelCardName": "string",
   "ModelCardStatus": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Content

The updated model card content. Content must be in model card JSON schema and provided as a string.

When updating model card content, be sure to include the full content and not just updated content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 100000.

Pattern: `.*`

Required: No

ModelCardName

The name or Amazon Resource Name (ARN) of the model card to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:model-card/.*)?([a-zA-Z0-9](-*[a-zA-Z0-9]){0,62})`

Required: Yes

ModelCardStatus

The approval status of the model card within your organization. Different organizations might have different criteria for model card review and approval.

- `Draft`: The model card is a work in progress.
- `PendingReview`: The model card is pending review.
- `Approved`: The model card is approved.
- `Archived`: The model card is archived. No more updates should be made to the model card, but it can still be exported.

Type: String

Valid Values: `Draft | PendingReview | Approved | Archived`

Required: No

Response Syntax

```
{
   "ModelCardArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelCardArn

The Amazon Resource Name (ARN) of the updated model card.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-card/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateModelPackage

Service: Amazon SageMaker Service

Updates a versioned model.

Request Syntax

```
{
   "AdditionalInferenceSpecificationsToAdd": [
      {
         "Containers": [
            {
               "AdditionalModelDataSources": [
                  {
                     "ChannelName": "string",
                     "S3DataSource": {
                        "CompressionType": "string",
                        "ETag": "string",
                        "HubAccessConfig": {
                           "HubContentArn": "string"
                        },
                        "ManifestEtag": "string",
                        "ManifestS3Uri": "string",
                        "ModelAccessConfig": {
                           "AcceptEula": boolean
                        },
                        "S3DataType": "string",
                        "S3Uri": "string"
                     }
                  }
               ],
               "AdditionalS3DataSource": {
                  "CompressionType": "string",
                  "ETag": "string",
                  "S3DataType": "string",
                  "S3Uri": "string"
               },
               "BaseModel": {
                  "HubContentName": "string",
                  "HubContentVersion": "string",
                  "RecipeName": "string"
               },
               "ContainerHostname": "string",
```

```
            "Environment": {
               "string" : "string"
            },
            "Framework": "string",
            "FrameworkVersion": "string",
            "Image": "string",
            "ImageDigest": "string",
            "IsCheckpoint": boolean,
            "ModelDataETag": "string",
            "ModelDataSource": {
               "S3DataSource": {
                  "CompressionType": "string",
                  "ETag": "string",
                  "HubAccessConfig": {
                     "HubContentArn": "string"
                  },
                  "ManifestEtag": "string",
                  "ManifestS3Uri": "string",
                  "ModelAccessConfig": {
                     "AcceptEula": boolean
                  },
                  "S3DataType": "string",
                  "S3Uri": "string"
               }
            },
            "ModelDataUrl": "string",
            "ModelInput": {
               "DataInputConfig": "string"
            },
            "NearestModelName": "string",
            "ProductId": "string"
         }
      ],
      "Description": "string",
      "Name": "string",
      "SupportedContentTypes": [ "string" ],
      "SupportedRealtimeInferenceInstanceTypes": [ "string" ],
      "SupportedResponseMIMETypes": [ "string" ],
      "SupportedTransformInstanceTypes": [ "string" ]
   }
],
"ApprovalDescription": "string",
"ClientToken": "string",
"CustomerMetadataProperties": {
```

```
      "string" : "string"
   },
   "CustomerMetadataPropertiesToRemove": [ "string" ],
   "InferenceSpecification": {
      "Containers": [
         {
            "AdditionalModelDataSources": [
               {
                  "ChannelName": "string",
                  "S3DataSource": {
                     "CompressionType": "string",
                     "ETag": "string",
                     "HubAccessConfig": {
                        "HubContentArn": "string"
                     },
                     "ManifestEtag": "string",
                     "ManifestS3Uri": "string",
                     "ModelAccessConfig": {
                        "AcceptEula": boolean
                     },
                     "S3DataType": "string",
                     "S3Uri": "string"
                  }
               }
            ],
            "AdditionalS3DataSource": {
               "CompressionType": "string",
               "ETag": "string",
               "S3DataType": "string",
               "S3Uri": "string"
            },
            "BaseModel": {
               "HubContentName": "string",
               "HubContentVersion": "string",
               "RecipeName": "string"
            },
            "ContainerHostname": "string",
            "Environment": {
               "string" : "string"
            },
            "Framework": "string",
            "FrameworkVersion": "string",
            "Image": "string",
            "ImageDigest": "string",
```

```
            "IsCheckpoint": boolean,
            "ModelDataETag": "string",
            "ModelDataSource": {
               "S3DataSource": {
                  "CompressionType": "string",
                  "ETag": "string",
                  "HubAccessConfig": {
                     "HubContentArn": "string"
                  },
                  "ManifestEtag": "string",
                  "ManifestS3Uri": "string",
                  "ModelAccessConfig": {
                     "AcceptEula": boolean
                  },
                  "S3DataType": "string",
                  "S3Uri": "string"
               }
            },
            "ModelDataUrl": "string",
            "ModelInput": {
               "DataInputConfig": "string"
            },
            "NearestModelName": "string",
            "ProductId": "string"
         }
      ],
      "SupportedContentTypes": [ "string" ],
      "SupportedRealtimeInferenceInstanceTypes": [ "string" ],
      "SupportedResponseMIMETypes": [ "string" ],
      "SupportedTransformInstanceTypes": [ "string" ]
   },
   "ModelApprovalStatus": "string",
   "ModelCard": {
      "ModelCardContent": "string",
      "ModelCardStatus": "string"
   },
   "ModelLifeCycle": {
      "Stage": "string",
      "StageDescription": "string",
      "StageStatus": "string"
   },
   "ModelPackageArn": "string",
   "ModelPackageRegistrationType": "string",
   "SourceUri": "string"
```

```
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AdditionalInferenceSpecificationsToAdd

An array of additional Inference Specification objects to be added to the existing array additional Inference Specification. Total number of additional Inference Specifications can not exceed 15. Each additional Inference Specification specifies artifacts based on this model package that can be used on inference endpoints. Generally used with SageMaker Neo to store the compiled artifacts.

Type: Array of AdditionalInferenceSpecificationDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 15 items.

Required: No

ApprovalDescription

A description for the approval status of the model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

ClientToken

A unique token that guarantees that the call to this API is idempotent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 36.

Pattern: `[a-zA-Z0-9-]+`

Required: No

CustomerMetadataProperties

The metadata properties associated with the model package versions.

Type: String to string map

Map Entries: Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `([\p{L}\p{Z}\p{N}_.:\/=+\-@]*)${1,128}`

Value Length Constraints: Minimum length of 1. Maximum length of 256.

Value Pattern: `([\p{L}\p{Z}\p{N}_.:\/=+\-@]*)${1,256}`

Required: No

CustomerMetadataPropertiesToRemove

The metadata properties associated with the model package versions to remove.

Type: Array of strings

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `([\p{L}\p{Z}\p{N}_.:\/=+\-@]*)${1,128}`

Required: No

InferenceSpecification

Specifies details about inference jobs that you can run with models based on this model package, including the following information:

- The Amazon ECR paths of containers that contain the inference code and model artifacts.
- The instance types that the model package supports for transform jobs and real-time endpoints used for inference.
- The input and output content formats that the model package supports for inference.

Type: InferenceSpecification object

Required: No

ModelApprovalStatus

The approval status of the model.

Type: String

Valid Values: `Approved | Rejected | PendingManualApproval`

Required: No

ModelCard

The model card associated with the model package. Since `ModelPackageModelCard` is tied to a model package, it is a specific usage of a model card and its schema is simplified compared to the schema of `ModelCard`. The `ModelPackageModelCard` schema does not include `model_package_details`, and `model_overview` is composed of the `model_creator` and `model_artifact` properties. For more information about the model package model card schema, see Model package model card schema. For more information about the model card associated with the model package, see View the Details of a Model Version.

Type: ModelPackageModelCard object

Required: No

ModelLifeCycle

A structure describing the current state of the model in its life cycle.

Type: ModelLifeCycle object

Required: No

ModelPackageArn

The Amazon Resource Name (ARN) of the model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: Yes

ModelPackageRegistrationType

The package registration type of the model package input.

Type: String

Valid Values: `Logged | Registered`

Required: No

SourceUri

The URI of the source for the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{N}\p{P}]{0,1024}`

Required: No

Response Syntax

```
{
   "ModelPackageArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ModelPackageArn

The Amazon Resource Name (ARN) of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateMonitoringAlert

Service: Amazon SageMaker Service

Update the parameters of a model monitor alert.

Request Syntax

```
{
   "DatapointsToAlert": number,
   "EvaluationPeriod": number,
   "MonitoringAlertName": "string",
   "MonitoringScheduleName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DatapointsToAlert

Within `EvaluationPeriod`, how many execution failures will raise an alert.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: Yes

EvaluationPeriod

The number of most recent monitoring executions to consider when evaluating alert status.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: Yes

MonitoringAlertName

The name of a monitoring alert.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

MonitoringScheduleName

The name of a monitoring schedule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "MonitoringAlertName": "string",
   "MonitoringScheduleArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

MonitoringAlertName

The name of a monitoring alert.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

MonitoringScheduleArn

The Amazon Resource Name (ARN) of the monitoring schedule.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateMonitoringSchedule

Service: Amazon SageMaker Service

Updates a previously created schedule.

Request Syntax

```
{
   "MonitoringScheduleConfig": {
      "MonitoringJobDefinition": {
         "BaselineConfig": {
            "BaseliningJobName": "string",
            "ConstraintsResource": {
               "S3Uri": "string"
            },
            "StatisticsResource": {
               "S3Uri": "string"
            }
         },
         "Environment": {
            "string" : "string"
         },
         "MonitoringAppSpecification": {
            "ContainerArguments": [ "string" ],
            "ContainerEntrypoint": [ "string" ],
            "ImageUri": "string",
            "PostAnalyticsProcessorSourceUri": "string",
            "RecordPreprocessorSourceUri": "string"
         },
         "MonitoringInputs": [
            {
               "BatchTransformInput": {
                  "DataCapturedDestinationS3Uri": "string",
                  "DatasetFormat": {
                     "Csv": {
                        "Header": boolean
                     },
                     "Json": {
                        "Line": boolean
                     },
                     "Parquet": {
                     }
                  },
```

```
                "EndTimeOffset": "string",
                "ExcludeFeaturesAttribute": "string",
                "FeaturesAttribute": "string",
                "InferenceAttribute": "string",
                "LocalPath": "string",
                "ProbabilityAttribute": "string",
                "ProbabilityThresholdAttribute": number,
                "S3DataDistributionType": "string",
                "S3InputMode": "string",
                "StartTimeOffset": "string"
            },
            "EndpointInput": {
                "EndpointName": "string",
                "EndTimeOffset": "string",
                "ExcludeFeaturesAttribute": "string",
                "FeaturesAttribute": "string",
                "InferenceAttribute": "string",
                "LocalPath": "string",
                "ProbabilityAttribute": "string",
                "ProbabilityThresholdAttribute": number,
                "S3DataDistributionType": "string",
                "S3InputMode": "string",
                "StartTimeOffset": "string"
            }
        }
    ],
    "MonitoringOutputConfig": {
        "KmsKeyId": "string",
        "MonitoringOutputs": [
            {
                "S3Output": {
                    "LocalPath": "string",
                    "S3UploadMode": "string",
                    "S3Uri": "string"
                }
            }
        ]
    },
    "MonitoringResources": {
        "ClusterConfig": {
            "InstanceCount": number,
            "InstanceType": "string",
            "VolumeKmsKeyId": "string",
            "VolumeSizeInGB": number
```

```
            }
         },
         "NetworkConfig": {
            "EnableInterContainerTrafficEncryption": boolean,
            "EnableNetworkIsolation": boolean,
            "VpcConfig": {
               "SecurityGroupIds": [ "string" ],
               "Subnets": [ "string" ]
            }
         },
         "RoleArn": "string",
         "StoppingCondition": {
            "MaxRuntimeInSeconds": number
         }
      },
      "MonitoringJobDefinitionName": "string",
      "MonitoringType": "string",
      "ScheduleConfig": {
         "DataAnalysisEndTime": "string",
         "DataAnalysisStartTime": "string",
         "ScheduleExpression": "string"
      }
   },
   "MonitoringScheduleName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

MonitoringScheduleConfig

The configuration object that specifies the monitoring schedule and defines the monitoring job.

Type: MonitoringScheduleConfig object

Required: Yes

MonitoringScheduleName

The name of the monitoring schedule. The name must be unique within an AWS Region within an AWS account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "MonitoringScheduleArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

MonitoringScheduleArn

The Amazon Resource Name (ARN) of the monitoring schedule.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateNotebookInstance

Service: Amazon SageMaker Service

Updates a notebook instance. NotebookInstance updates include upgrading or downgrading the ML compute instance used for your notebook instance to accommodate changes in your workload requirements.

> **Note**
>
> This API can attach lifecycle configurations to notebook instances. Lifecycle configuration scripts execute with root access and the notebook instance's IAM execution role privileges. Principals with this permission and access to lifecycle configurations can execute code with the execution role's credentials. See Customize a Notebook Instance Using a Lifecycle Configuration Script for security best practices.

Request Syntax

```
{
   "AcceleratorTypes": [ "string" ],
   "AdditionalCodeRepositories": [ "string" ],
   "DefaultCodeRepository": "string",
   "DisassociateAcceleratorTypes": boolean,
   "DisassociateAdditionalCodeRepositories": boolean,
   "DisassociateDefaultCodeRepository": boolean,
   "DisassociateLifecycleConfig": boolean,
   "InstanceMetadataServiceConfiguration": {
      "MinimumInstanceMetadataServiceVersion": "string"
   },
   "InstanceType": "string",
   "IpAddressType": "string",
   "LifecycleConfigName": "string",
   "NotebookInstanceName": "string",
   "PlatformIdentifier": "string",
   "RoleArn": "string",
   "RootAccess": "string",
   "VolumeSizeInGB": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

AcceleratorTypes

This parameter is no longer supported. Elastic Inference (EI) is no longer available.

This parameter was used to specify a list of the EI instance types to associate with this notebook instance.

Type: Array of strings

Valid Values: `ml.eia1.medium | ml.eia1.large | ml.eia1.xlarge | ml.eia2.medium | ml.eia2.large | ml.eia2.xlarge`

Required: No

AdditionalCodeRepositories

An array of up to three Git repositories to associate with the notebook instance. These can be either the names of Git repositories stored as resources in your account, or the URL of Git repositories in AWS CodeCommit or in any other Git repository. These repositories are cloned at the same level as the default repository of your notebook instance. For more information, see Associating Git Repositories with SageMaker AI Notebook Instances.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 3 items.

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `https://([^/]+)/?(.*)$|^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

DefaultCodeRepository

The Git repository to associate with the notebook instance as its default code repository. This can be either the name of a Git repository stored as a resource in your account, or the URL of a Git repository in AWS CodeCommit or in any other Git repository. When you open a notebook

instance, it opens in the directory that contains this repository. For more information, see Associating Git Repositories with SageMaker AI Notebook Instances.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `https://([^/]+)/?(.*)$|^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

DisassociateAcceleratorTypes

This parameter is no longer supported. Elastic Inference (EI) is no longer available.

This parameter was used to specify a list of the EI instance types to remove from this notebook instance.

Type: Boolean

Required: No

DisassociateAdditionalCodeRepositories

A list of names or URLs of the default Git repositories to remove from this notebook instance. This operation is idempotent. If you specify a Git repository that is not associated with the notebook instance when you call this method, it does not throw an error.

Type: Boolean

Required: No

DisassociateDefaultCodeRepository

The name or URL of the default Git repository to remove from this notebook instance. This operation is idempotent. If you specify a Git repository that is not associated with the notebook instance when you call this method, it does not throw an error.

Type: Boolean

Required: No

DisassociateLifecycleConfig

Set to `true` to remove the notebook instance lifecycle configuration currently associated with the notebook instance. This operation is idempotent. If you specify a lifecycle configuration that

is not associated with the notebook instance when you call this method, it does not throw an error.

Type: Boolean

Required: No

InstanceMetadataServiceConfiguration

Information on the IMDS configuration of the notebook instance

Type: InstanceMetadataServiceConfiguration object

Required: No

InstanceType

The Amazon ML compute instance type.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.8xlarge | ml.m5d.12xlarge | ml.m5d.16xlarge | ml.m5d.24xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p3dn.24xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge`

```
| ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p6-
b200.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge |
ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge
| ml.m6i.24xlarge | ml.m6i.32xlarge | ml.m7i.large | ml.m7i.xlarge
| ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge
| ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c6i.large
| ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge |
ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge
| ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge |
ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge
| ml.c7i.48xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge |
ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge
| ml.r6i.24xlarge | ml.r6i.32xlarge | ml.r7i.large | ml.r7i.xlarge |
ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge |
ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.m6id.large |
ml.m6id.xlarge | ml.m6id.2xlarge | ml.m6id.4xlarge | ml.m6id.8xlarge
| ml.m6id.12xlarge | ml.m6id.16xlarge | ml.m6id.24xlarge |
ml.m6id.32xlarge | ml.c6id.large | ml.c6id.xlarge | ml.c6id.2xlarge |
ml.c6id.4xlarge | ml.c6id.8xlarge | ml.c6id.12xlarge | ml.c6id.16xlarge
| ml.c6id.24xlarge | ml.c6id.32xlarge | ml.r6id.large | ml.r6id.xlarge
| ml.r6id.2xlarge | ml.r6id.4xlarge | ml.r6id.8xlarge | ml.r6id.12xlarge
| ml.r6id.16xlarge | ml.r6id.24xlarge | ml.r6id.32xlarge | ml.g6.xlarge
| ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge |
ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.p5.4xlarge |
ml.p5en.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge |
ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge |
ml.g6e.48xlarge
```

Required: No

IpAddressType

The IP address type for the notebook instance. Specify `ipv4` for IPv4-only connectivity or `dualstack` for both IPv4 and IPv6 connectivity. The notebook instance must be stopped before updating this setting. When you specify `dualstack`, the subnet must support IPv6 addressing.

Type: String

Valid Values: `ipv4 | dualstack`

Required: No

LifecycleConfigName

The name of a lifecycle configuration to associate with the notebook instance. For information about lifestyle configurations, see Step 2.1: (Optional) Customize a Notebook Instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

NotebookInstanceName

The name of the notebook instance to update.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

PlatformIdentifier

The platform identifier of the notebook instance runtime environment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 20.

Pattern: `(notebook-al1-v1|notebook-al2-v1|notebook-al2-v2|notebook-al2-v3|notebook-al2023-v1)`

Required: No

RoleArn

The Amazon Resource Name (ARN) of the IAM role that SageMaker AI can assume to access the notebook instance. For more information, see SageMaker AI Roles.

> **Note**
>
> To be able to pass this role to SageMaker AI, the caller of this API must have the `iam:PassRole` permission.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

RootAccess

Whether root access is enabled or disabled for users of the notebook instance. The default value is `Enabled`.

> **Note**
>
> If you set this to `Disabled`, users don't have root access on the notebook instance, but lifecycle configuration scripts still run with root permissions.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

VolumeSizeInGB

The size, in GB, of the ML storage volume to attach to the notebook instance. The default value is 5 GB. ML storage volumes are encrypted, so SageMaker AI can't determine the amount of available free space on the volume. Because of this, you can increase the volume size when you update a notebook instance, but you can't decrease the volume size. If you want to decrease the size of the ML storage volume in use, create a new notebook instance with the desired size.

Type: Integer

Valid Range: Minimum value of 5. Maximum value of 16384.

Required: No

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateNotebookInstanceLifecycleConfig

Service: Amazon SageMaker Service

Updates a notebook instance lifecycle configuration created with the CreateNotebookInstanceLifecycleConfig API.

> **Note**
>
> Updates to lifecycle configurations affect all notebook instances using that configuration upon their next start. Lifecycle configuration scripts execute with root access and the notebook instance's IAM execution role privileges. Grant this permission only to trusted principals. See Customize a Notebook Instance Using a Lifecycle Configuration Script for security best practices.

Request Syntax

```
{
   "NotebookInstanceLifecycleConfigName": "string",
   "OnCreate": [
      {
         "Content": "string"
      }
   ],
   "OnStart": [
      {
         "Content": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

NotebookInstanceLifecycleConfigName

The name of the lifecycle configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

OnCreate

The shell script that runs only once, when you create a notebook instance. The shell script must be a base64-encoded string.

Type: Array of NotebookInstanceLifecycleHook objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

OnStart

The shell script that runs every time you start a notebook instance, including when you create the notebook instance. The shell script must be a base64-encoded string.

Type: Array of NotebookInstanceLifecycleHook objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdatePartnerApp

Service: Amazon SageMaker Service

Updates all of the SageMaker Partner AI Apps in an account.

Request Syntax

```
{
   "ApplicationConfig": {
      "AdminUsers": [ "string" ],
      "Arguments": {
         "string" : "string"
      },
      "AssignedGroupPatterns": [ "string" ],
      "RoleGroupAssignments": [
         {
            "GroupPatterns": [ "string" ],
            "RoleName": "string"
         }
      ]
   },
   "AppVersion": "string",
   "Arn": "string",
   "ClientToken": "string",
   "EnableAutoMinorVersionUpgrade": boolean,
   "EnableIamSessionBasedIdentity": boolean,
   "MaintenanceConfig": {
      "MaintenanceWindowStart": "string"
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "Tier": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ApplicationConfig

Configuration settings for the SageMaker Partner AI App.

Type: PartnerAppConfig object

Required: No

AppVersion

The semantic version to upgrade the SageMaker Partner AI App to. Must be the same semantic version returned in the `AvailableUpgrade` field from `DescribePartnerApp`. Version skipping and downgrades are not supported.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `\d+\.\d+`

Required: No

Arn

The ARN of the SageMaker Partner AI App to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:partner-app\/app-[A-Z0-9]{12}`

Required: Yes

ClientToken

A unique token that guarantees that the call to this API is idempotent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 36.

Pattern: `[a-zA-Z0-9-]+`

Required: No

EnableAutoMinorVersionUpgrade

When set to TRUE, the SageMaker Partner AI App is automatically upgraded to the latest minor version during the next scheduled maintenance window, if one is available.

Type: Boolean

Required: No

EnableIamSessionBasedIdentity

When set to TRUE, the SageMaker Partner AI App sets the AWS IAM session name or the authenticated IAM user as the identity of the SageMaker Partner AI App user.

Type: Boolean

Required: No

MaintenanceConfig

Maintenance configuration settings for the SageMaker Partner AI App.

Type: PartnerAppMaintenanceConfig object

Required: No

Tags

Each tag consists of a key and an optional value. Tag keys must be unique per resource.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Tier

Indicates the instance type and size of the cluster attached to the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `(?!\s*$).+`

Required: No

Response Syntax

```
{
   "Arn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The ARN of the SageMaker Partner AI App that was updated.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:partner-app\/app-[A-Z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdatePipeline

Service: Amazon SageMaker Service

Updates a pipeline.

Request Syntax

```
{
   "ParallelismConfiguration": {
      "MaxParallelExecutionSteps": number
   },
   "PipelineDefinition": "string",
   "PipelineDefinitionS3Location": {
      "Bucket": "string",
      "ObjectKey": "string",
      "VersionId": "string"
   },
   "PipelineDescription": "string",
   "PipelineDisplayName": "string",
   "PipelineName": "string",
   "RoleArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ParallelismConfiguration

If specified, it applies to all executions of this pipeline by default.

Type: ParallelismConfiguration object

Required: No

PipelineDefinition

The JSON pipeline definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1048576.

Pattern: `.*(?:[ \r\n\t].*)*`

Required: No

PipelineDefinitionS3Location

The location of the pipeline definition stored in Amazon S3. If specified, SageMaker will retrieve the pipeline definition from this location.

Type: PipelineDefinitionS3Location object

Required: No

PipelineDescription

The description of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

PipelineDisplayName

The display name of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: No

PipelineName

The name of the pipeline to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: Yes

RoleArn

The Amazon Resource Name (ARN) that the pipeline uses to execute.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

Response Syntax

```
{
   "PipelineArn": "string",
   "PipelineVersionId": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

PipelineArn

The Amazon Resource Name (ARN) of the updated pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

PipelineVersionId

The ID of the pipeline version.

Type: Long

Valid Range: Minimum value of 1.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdatePipelineExecution

Service: Amazon SageMaker Service

Updates a pipeline execution.

Request Syntax

```
{
   "ParallelismConfiguration": {
      "MaxParallelExecutionSteps": number
   },
   "PipelineExecutionArn": "string",
   "PipelineExecutionDescription": "string",
   "PipelineExecutionDisplayName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ParallelismConfiguration

This configuration, if specified, overrides the parallelism configuration of the parent pipeline for this specific run.

Type: ParallelismConfiguration object

Required: No

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: Yes

PipelineExecutionDescription

The description of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

PipelineExecutionDisplayName

The display name of the pipeline execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 82.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,81}`

Required: No

Response Syntax

```
{
   "PipelineExecutionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

PipelineExecutionArn

The Amazon Resource Name (ARN) of the updated pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdatePipelineVersion

Service: Amazon SageMaker Service

Updates a pipeline version.

Request Syntax

```
{
   "PipelineArn": "string",
   "PipelineVersionDescription": "string",
   "PipelineVersionDisplayName": "string",
   "PipelineVersionId": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

PipelineArn

The Amazon Resource Name (ARN) of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

Required: Yes

PipelineVersionDescription

The description of the pipeline version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

PipelineVersionDisplayName

The display name of the pipeline version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 82.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,81}`

Required: No

PipelineVersionId

The pipeline version ID to update.

Type: Long

Valid Range: Minimum value of 1.

Required: Yes

Response Syntax

```
{
   "PipelineArn": "string",
   "PipelineVersionId": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

PipelineArn

The Amazon Resource Name (ARN) of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

PipelineVersionId

The ID of the pipeline version.

Type: Long

Valid Range: Minimum value of 1.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateProject

Service: Amazon SageMaker Service

Updates a machine learning (ML) project that is created from a template that sets up an ML pipeline from training to deploying an approved model.

> **Note**
>
> You must not update a project that is in use. If you update the `ServiceCatalogProvisioningUpdateDetails` of a project that is active or being created, or updated, you may lose resources already created by the project.

Request Syntax

```
{
   "ProjectDescription": "string",
   "ProjectName": "string",
   "ServiceCatalogProvisioningUpdateDetails": {
      "ProvisioningArtifactId": "string",
      "ProvisioningParameters": [
         {
            "Key": "string",
            "Value": "string"
         }
      ]
   },
   "Tags": [
      {
         "Key": "string",
         "Value": "string"
      }
   ],
   "TemplateProvidersToUpdate": [
      {
         "CfnTemplateProvider": {
            "Parameters": [
               {
                  "Key": "string",
                  "Value": "string"
               }
            ],
```

```
            "TemplateName": "string",
            "TemplateURL": "string"
         }
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ProjectDescription

The description for the project.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

ProjectName

The name of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

ServiceCatalogProvisioningUpdateDetails

The product ID and provisioning artifact ID to provision a service catalog. The provisioning artifact ID will default to the latest provisioning artifact ID of the product, if you don't provide the provisioning artifact ID. For more information, see What is AWS Service Catalog.

Type: ServiceCatalogProvisioningUpdateDetails object

Required: No

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources. In addition, the project must have tag update constraints set in order to include this parameter in the request. For more information, see AWS Service Catalog Tag Update Constraints.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TemplateProvidersToUpdate

The template providers to update in the project.

Type: Array of UpdateTemplateProvider objects

Array Members: Fixed number of 1 item.

Required: No

Response Syntax

```
{
   "ProjectArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ProjectArn

The Amazon Resource Name (ARN) of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:project/[\S]{1,2048}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateSpace

Service: Amazon SageMaker Service

Updates the settings of a space.

> **Note**
>
> You can't edit the app type of a space in the `SpaceSettings`.

Request Syntax

```
{
   "DomainId": "string",
   "SpaceDisplayName": "string",
   "SpaceName": "string",
   "SpaceSettings": {
      "AppType": "string",
      "CodeEditorAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number
            }
         },
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         }
      },
      "CustomFileSystems": [
         { ... }
      ],
      "JupyterLabAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number
            }
         },
```

```
      "CodeRepositories": [
         {
            "RepositoryUrl": "string"
         }
      ],
      "DefaultResourceSpec": {
         "InstanceType": "string",
         "LifecycleConfigArn": "string",
         "SageMakerImageArn": "string",
         "SageMakerImageVersionAlias": "string",
         "SageMakerImageVersionArn": "string",
         "TrainingPlanArn": "string"
      }
   },
   "JupyterServerAppSettings": {
      "CodeRepositories": [
         {
            "RepositoryUrl": "string"
         }
      ],
      "DefaultResourceSpec": {
         "InstanceType": "string",
         "LifecycleConfigArn": "string",
         "SageMakerImageArn": "string",
         "SageMakerImageVersionAlias": "string",
         "SageMakerImageVersionArn": "string",
         "TrainingPlanArn": "string"
      },
      "LifecycleConfigArns": [ "string" ]
   },
   "KernelGatewayAppSettings": {
      "CustomImages": [
         {
            "AppImageConfigName": "string",
            "ImageName": "string",
            "ImageVersionNumber": number
         }
      ],
      "DefaultResourceSpec": {
         "InstanceType": "string",
         "LifecycleConfigArn": "string",
         "SageMakerImageArn": "string",
         "SageMakerImageVersionAlias": "string",
         "SageMakerImageVersionArn": "string",
```

```
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "RemoteAccess": "string",
      "SpaceManagedResources": "string",
      "SpaceStorageSettings": {
         "EbsStorageSettings": {
            "EbsVolumeSizeInGb": number
         }
      }
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainId

The ID of the associated domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

SpaceDisplayName

The name of the space that appears in the Amazon SageMaker Studio UI.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `(?!\s*$).+`

Required: No

SpaceName

The name of the space.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

SpaceSettings

A collection of space settings.

Type: SpaceSettings object

Required: No

Response Syntax

```
{
   "SpaceArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

SpaceArn

The space's Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:space/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateTrainingJob

Service: Amazon SageMaker Service

Update a model training job to request a new Debugger profiling configuration or to change warm pool retention length.

Request Syntax

```
{
   "ProfilerConfig": {
      "DisableProfiler": boolean,
      "ProfilingIntervalInMilliseconds": number,
      "ProfilingParameters": {
         "string" : "string"
      },
      "S3OutputPath": "string"
   },
   "ProfilerRuleConfigurations": [
      {
         "InstanceType": "string",
         "LocalPath": "string",
         "RuleConfigurationName": "string",
         "RuleEvaluatorImage": "string",
         "RuleParameters": {
            "string" : "string"
         },
         "S3OutputPath": "string",
         "VolumeSizeInGB": number
      }
   ],
   "RemoteDebugConfig": {
      "EnableRemoteDebug": boolean
   },
   "ResourceConfig": {
      "KeepAlivePeriodInSeconds": number
   },
   "TrainingJobName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

ProfilerConfig

Configuration information for Amazon SageMaker Debugger system monitoring, framework profiling, and storage paths.

Type: ProfilerConfigForUpdate object

Required: No

ProfilerRuleConfigurations

Configuration information for Amazon SageMaker Debugger rules for profiling system and framework metrics.

Type: Array of ProfilerRuleConfiguration objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

RemoteDebugConfig

Configuration for remote debugging while the training job is running. You can update the remote debugging configuration when the `SecondaryStatus` of the job is `Downloading` or `Training`.To learn more about the remote debugging functionality of SageMaker, see Access a training container through AWS Systems Manager (SSM) for remote debugging.

Type: RemoteDebugConfigForUpdate object

Required: No

ResourceConfig

The training job `ResourceConfig` to update warm pool retention length.

Type: ResourceConfigForUpdate object

Required: No

TrainingJobName

The name of a training job to update the Debugger profiling configuration.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "TrainingJobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrainingJobArn

The Amazon Resource Name (ARN) of the training job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateTrial

Service: Amazon SageMaker Service

Updates the display name of a trial.

Request Syntax

```
{
   "DisplayName": "string",
   "TrialName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DisplayName

The name of the trial as displayed. The name doesn't need to be unique. If `DisplayName` isn't specified, `TrialName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

TrialName

The name of the trial to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
   "TrialArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrialArn

The Amazon Resource Name (ARN) of the trial.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateTrialComponent

Service: Amazon SageMaker Service

Updates one or more properties of a trial component.

Request Syntax

```
{
   "DisplayName": "string",
   "EndTime": number,
   "InputArtifacts": {
      "string" : {
         "MediaType": "string",
         "Value": "string"
      }
   },
   "InputArtifactsToRemove": [ "string" ],
   "OutputArtifacts": {
      "string" : {
         "MediaType": "string",
         "Value": "string"
      }
   },
   "OutputArtifactsToRemove": [ "string" ],
   "Parameters": {
      "string" : {
         "NumberValue": number,
         "StringValue": "string"
      }
   },
   "ParametersToRemove": [ "string" ],
   "StartTime": number,
   "Status": {
      "Message": "string",
      "PrimaryStatus": "string"
   },
   "TrialComponentName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DisplayName

The name of the component as displayed. The name doesn't need to be unique. If `DisplayName` isn't specified, `TrialComponentName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

EndTime

When the component ended.

Type: Timestamp

Required: No

InputArtifacts

Replaces all of the component's input artifacts with the specified artifacts or adds new input artifacts. Existing input artifacts are replaced if the trial component is updated with an identical input artifact key.

Type: String to TrialComponentArtifact object map

Map Entries: Minimum number of 0 items. Maximum number of 60 items.

Key Length Constraints: Minimum length of 0. Maximum length of 128.

Key Pattern: `.*`

Required: No

InputArtifactsToRemove

The input artifacts to remove from the component.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

OutputArtifacts

Replaces all of the component's output artifacts with the specified artifacts or adds new output artifacts. Existing output artifacts are replaced if the trial component is updated with an identical output artifact key.

Type: String to TrialComponentArtifact object map

Map Entries: Minimum number of 0 items. Maximum number of 60 items.

Key Length Constraints: Minimum length of 0. Maximum length of 128.

Key Pattern: `.*`

Required: No

OutputArtifactsToRemove

The output artifacts to remove from the component.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

Parameters

Replaces all of the component's hyperparameters with the specified hyperparameters or add new hyperparameters. Existing hyperparameters are replaced if the trial component is updated with an identical hyperparameter key.

Type: String to TrialComponentParameterValue object map

Map Entries: Minimum number of 0 items. Maximum number of 300 items.

Key Length Constraints: Minimum length of 0. Maximum length of 320.

Key Pattern: `.*`

Required: No

ParametersToRemove

The hyperparameters to remove from the component.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

StartTime

When the component started.

Type: Timestamp

Required: No

Status

The new status of the component.

Type: TrialComponentStatus object

Required: No

TrialComponentName

The name of the component to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Response Syntax

```
{
```

```
   "TrialComponentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

TrialComponentArn

The Amazon Resource Name (ARN) of the trial component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial-component/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateUserProfile

Service: Amazon SageMaker Service

Updates a user profile.

Request Syntax

```
{
   "DomainId": "string",
   "UserProfileName": "string",
   "UserSettings": {
      "AutoMountHomeEFS": "string",
      "CanvasAppSettings": {
         "DirectDeploySettings": {
            "Status": "string"
         },
         "EmrServerlessSettings": {
            "ExecutionRoleArn": "string",
            "Status": "string"
         },
         "GenerativeAiSettings": {
            "AmazonBedrockRoleArn": "string"
         },
         "IdentityProviderOAuthSettings": [
            {
               "DataSourceName": "string",
               "SecretArn": "string",
               "Status": "string"
            }
         ],
         "KendraSettings": {
            "Status": "string"
         },
         "ModelRegisterSettings": {
            "CrossAccountModelRegisterRoleArn": "string",
            "Status": "string"
         },
         "TimeSeriesForecastingSettings": {
            "AmazonForecastRoleArn": "string",
            "Status": "string"
         },
         "WorkspaceSettings": {
            "S3ArtifactPath": "string",
```

```
            "S3KmsKeyId": "string"
         }
      },
      "CodeEditorAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number,
               "LifecycleManagement": "string",
               "MaxIdleTimeoutInMinutes": number,
               "MinIdleTimeoutInMinutes": number
            }
         },
         "BuiltInLifecycleConfigArn": "string",
         "CustomImages": [
            {
               "AppImageConfigName": "string",
               "ImageName": "string",
               "ImageVersionNumber": number
            }
         ],
         "DefaultResourceSpec": {
            "InstanceType": "string",
            "LifecycleConfigArn": "string",
            "SageMakerImageArn": "string",
            "SageMakerImageVersionAlias": "string",
            "SageMakerImageVersionArn": "string",
            "TrainingPlanArn": "string"
         },
         "LifecycleConfigArns": [ "string" ]
      },
      "CustomFileSystemConfigs": [
         { ... }
      ],
      "CustomPosixUserConfig": {
         "Gid": number,
         "Uid": number
      },
      "DefaultLandingUri": "string",
      "ExecutionRole": "string",
      "JupyterLabAppSettings": {
         "AppLifecycleManagement": {
            "IdleSettings": {
               "IdleTimeoutInMinutes": number,
               "LifecycleManagement": "string",
```

```
            "MaxIdleTimeoutInMinutes": number,
            "MinIdleTimeoutInMinutes": number
         }
      },
      "BuiltInLifecycleConfigArn": "string",
      "CodeRepositories": [
         {
            "RepositoryUrl": "string"
         }
      ],
      "CustomImages": [
         {
            "AppImageConfigName": "string",
            "ImageName": "string",
            "ImageVersionNumber": number
         }
      ],
      "DefaultResourceSpec": {
         "InstanceType": "string",
         "LifecycleConfigArn": "string",
         "SageMakerImageArn": "string",
         "SageMakerImageVersionAlias": "string",
         "SageMakerImageVersionArn": "string",
         "TrainingPlanArn": "string"
      },
      "EmrSettings": {
         "AssumableRoleArns": [ "string" ],
         "ExecutionRoleArns": [ "string" ]
      },
      "LifecycleConfigArns": [ "string" ]
   },
   "JupyterServerAppSettings": {
      "CodeRepositories": [
         {
            "RepositoryUrl": "string"
         }
      ],
      "DefaultResourceSpec": {
         "InstanceType": "string",
         "LifecycleConfigArn": "string",
         "SageMakerImageArn": "string",
         "SageMakerImageVersionAlias": "string",
         "SageMakerImageVersionArn": "string",
         "TrainingPlanArn": "string"
```

```
            },
            "LifecycleConfigArns": [ "string" ]
         },
         "KernelGatewayAppSettings": {
            "CustomImages": [
               {
                  "AppImageConfigName": "string",
                  "ImageName": "string",
                  "ImageVersionNumber": number
               }
            ],
            "DefaultResourceSpec": {
               "InstanceType": "string",
               "LifecycleConfigArn": "string",
               "SageMakerImageArn": "string",
               "SageMakerImageVersionAlias": "string",
               "SageMakerImageVersionArn": "string",
               "TrainingPlanArn": "string"
            },
            "LifecycleConfigArns": [ "string" ]
         },
         "RSessionAppSettings": {
            "CustomImages": [
               {
                  "AppImageConfigName": "string",
                  "ImageName": "string",
                  "ImageVersionNumber": number
               }
            ],
            "DefaultResourceSpec": {
               "InstanceType": "string",
               "LifecycleConfigArn": "string",
               "SageMakerImageArn": "string",
               "SageMakerImageVersionAlias": "string",
               "SageMakerImageVersionArn": "string",
               "TrainingPlanArn": "string"
            }
         },
         "RStudioServerProAppSettings": {
            "AccessStatus": "string",
            "UserGroup": "string"
         },
         "SecurityGroups": [ "string" ],
         "SharingSettings": {
```

```
            "NotebookOutputOption": "string",
            "S3KmsKeyId": "string",
            "S3OutputPath": "string"
         },
         "SpaceStorageSettings": {
            "DefaultEbsStorageSettings": {
               "DefaultEbsVolumeSizeInGb": number,
               "MaximumEbsVolumeSizeInGb": number
            }
         },
         "StudioWebPortal": "string",
         "StudioWebPortalSettings": {
            "ExecutionRoleSessionNameMode": "string",
            "HiddenAppTypes": [ "string" ],
            "HiddenInstanceTypes": [ "string" ],
            "HiddenMlTools": [ "string" ],
            "HiddenSageMakerImageVersionAliases": [
               {
                  "SageMakerImageName": "string",
                  "VersionAliases": [ "string" ]
               }
            ]
         },
         "TensorBoardAppSettings": {
            "DefaultResourceSpec": {
               "InstanceType": "string",
               "LifecycleConfigArn": "string",
               "SageMakerImageArn": "string",
               "SageMakerImageVersionAlias": "string",
               "SageMakerImageVersionArn": "string",
               "TrainingPlanArn": "string"
            }
         }
      }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: Yes

UserProfileName

The user profile name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

UserSettings

A collection of settings.

Type: UserSettings object

Required: No

Response Syntax

```
{
   "UserProfileArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

UserProfileArn

The user profile Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:user-profile/.*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceInUse

Resource being accessed is in use.

HTTP Status Code: 400

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

ResourceNotFound

Resource being access is not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateWorkforce

Service: Amazon SageMaker Service

Use this operation to update your workforce. You can use this operation to require that workers use specific IP addresses to work on tasks and to update your OpenID Connect (OIDC) Identity Provider (IdP) workforce configuration.

The worker portal is now supported in VPC and public internet.

Use `SourceIpConfig` to restrict worker access to tasks to a specific range of IP addresses. You specify allowed IP addresses by creating a list of up to ten CIDRs. By default, a workforce isn't restricted to specific IP addresses. If you specify a range of IP addresses, workers who attempt to access tasks using any IP address outside the specified range are denied and get a `Not Found` error message on the worker portal.

To restrict public internet access for all workers, configure the `SourceIpConfig` CIDR value. For example, when using `SourceIpConfig` with an `IpAddressType` of `IPv4`, you can restrict access to the IPv4 CIDR block "10.0.0.0/16". When using an `IpAddressType` of `dualstack`, you can specify both the IPv4 and IPv6 CIDR blocks, such as "10.0.0.0/16" for IPv4 only, "2001:db8:1234:1a00::/56" for IPv6 only, or "10.0.0.0/16" and "2001:db8:1234:1a00::/56" for dual stack.

> **Important**
>
> Amazon SageMaker does not support Source Ip restriction for worker portals in VPC.

Use `OidcConfig` to update the configuration of a workforce created using your own OIDC IdP.

> **Important**
>
> You can only update your OIDC IdP configuration when there are no work teams associated with your workforce. You can delete work teams using the DeleteWorkteam operation.

After restricting access to a range of IP addresses or updating your OIDC IdP configuration with this operation, you can view details about your update workforce using the DescribeWorkforce operation.

> **Important**
>
> This operation only applies to private workforces.

Request Syntax

```
{
   "IpAddressType": "string",
   "OidcConfig": {
      "AuthenticationRequestExtraParams": {
         "string" : "string"
      },
      "AuthorizationEndpoint": "string",
      "ClientId": "string",
      "ClientSecret": "string",
      "Issuer": "string",
      "JwksUri": "string",
      "LogoutEndpoint": "string",
      "Scope": "string",
      "TokenEndpoint": "string",
      "UserInfoEndpoint": "string"
   },
   "SourceIpConfig": {
      "Cidrs": [ "string" ]
   },
   "WorkforceName": "string",
   "WorkforceVpcConfig": {
      "SecurityGroupIds": [ "string" ],
      "Subnets": [ "string" ],
      "VpcId": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

IpAddressType

Use this parameter to specify whether you want `IPv4` only or `dualstack` (`IPv4` and `IPv6`) to support your labeling workforce.

Type: String

Valid Values: `ipv4 | dualstack`

Required: No

OidcConfig

Use this parameter to update your OIDC Identity Provider (IdP) configuration for a workforce made using your own IdP.

Type: OidcConfig object

Required: No

SourceIpConfig

A list of one to ten worker IP address ranges (CIDRs) that can be used to access tasks assigned to this workforce.

Maximum: Ten CIDR values

Type: SourceIpConfig object

Required: No

WorkforceName

The name of the private workforce that you want to update. You can find your workforce name by using the ListWorkforces operation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9\-]){0,62}`

Required: Yes

WorkforceVpcConfig

Use this parameter to update your VPC configuration for a workforce.

Type: WorkforceVpcConfigRequest object

Required: No

Response Syntax

```
{
   "Workforce": {
      "CognitoConfig": {
         "ClientId": "string",
         "UserPool": "string"
      },
      "CreateDate": number,
      "FailureReason": "string",
      "IpAddressType": "string",
      "LastUpdatedDate": number,
      "OidcConfig": {
         "AuthenticationRequestExtraParams": {
            "string" : "string"
         },
         "AuthorizationEndpoint": "string",
         "ClientId": "string",
         "Issuer": "string",
         "JwksUri": "string",
         "LogoutEndpoint": "string",
         "Scope": "string",
         "TokenEndpoint": "string",
         "UserInfoEndpoint": "string"
      },
      "SourceIpConfig": {
         "Cidrs": [ "string" ]
      },
      "Status": "string",
      "SubDomain": "string",
      "WorkforceArn": "string",
      "WorkforceName": "string",
      "WorkforceVpcConfig": {
         "SecurityGroupIds": [ "string" ],
         "Subnets": [ "string" ],
         "VpcEndpointId": "string",
         "VpcId": "string"
      }
   }
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Workforce

A single private workforce. You can create one private work force in each AWS Region. By default, any workforce-related API operation used in a specific region will apply to the workforce created in that region. To learn how to create a private workforce, see Create a Private Workforce.

Type: Workforce object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ConflictException

There was a conflict when you attempted to modify a SageMaker entity such as an `Experiment` or `Artifact`.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateWorkteam

Service: Amazon SageMaker Service

Updates an existing work team with new member definitions or description.

Request Syntax

```
{
   "Description": "string",
   "MemberDefinitions": [
      {
         "CognitoMemberDefinition": {
            "ClientId": "string",
            "UserGroup": "string",
            "UserPool": "string"
         },
         "OidcMemberDefinition": {
            "Groups": [ "string" ]
         }
      }
   ],
   "NotificationConfiguration": {
      "NotificationTopicArn": "string"
   },
   "WorkerAccessConfiguration": {
      "S3Presign": {
         "IamPolicyConstraints": {
            "SourceIp": "string",
            "VpcSourceIp": "string"
         }
      }
   },
   "WorkteamName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

Description

An updated description for the work team.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `.+`

Required: No

MemberDefinitions

A list of `MemberDefinition` objects that contains objects that identify the workers that make up the work team.

Workforces can be created using Amazon Cognito or your own OIDC Identity Provider (IdP). For private workforces created using Amazon Cognito use `CognitoMemberDefinition`. For workforces created using your own OIDC identity provider (IdP) use `OidcMemberDefinition`. You should not provide input for both of these parameters in a single request.

For workforces created using Amazon Cognito, private work teams correspond to Amazon Cognito *user groups* within the user pool used to create a workforce. All of the `CognitoMemberDefinition` objects that make up the member definition must have the same `ClientId` and `UserPool` values. To add a Amazon Cognito user group to an existing worker pool, see Adding groups to a User Pool. For more information about user pools, see Amazon Cognito User Pools.

For workforces created using your own OIDC IdP, specify the user groups that you want to include in your private work team in `OidcMemberDefinition` by listing those groups in `Groups`. Be aware that user groups that are already in the work team must also be listed in `Groups` when you make this request to remain on the work team. If you do not include these user groups, they will no longer be associated with the work team you update.

Type: Array of MemberDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

NotificationConfiguration

Configures SNS topic notifications for available or expiring work items

Type: NotificationConfiguration object

Required: No

WorkerAccessConfiguration

Use this optional parameter to constrain access to an Amazon S3 resource based on the IP address using supported IAM global condition keys. The Amazon S3 resource is accessed in the worker portal using a Amazon S3 presigned URL.

Type: WorkerAccessConfiguration object

Required: No

WorkteamName

The name of the work team to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Response Syntax

```
{
   "Workteam": {
      "CreateDate": number,
      "Description": "string",
      "LastUpdatedDate": number,
      "MemberDefinitions": [
         {
            "CognitoMemberDefinition": {
               "ClientId": "string",
               "UserGroup": "string",
               "UserPool": "string"
            },
            "OidcMemberDefinition": {
               "Groups": [ "string" ]
            }
         }
```

```
      ],
      "NotificationConfiguration": {
         "NotificationTopicArn": "string"
      },
      "ProductListingIds": [ "string" ],
      "SubDomain": "string",
      "WorkerAccessConfiguration": {
         "S3Presign": {
            "IamPolicyConstraints": {
               "SourceIp": "string",
               "VpcSourceIp": "string"
            }
         }
      },
      "WorkforceArn": "string",
      "WorkteamArn": "string",
      "WorkteamName": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Workteam

A `Workteam` object that describes the updated work team.

Type: Workteam object

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceLimitExceeded

You have exceeded an SageMaker resource limit. For example, you might have too many training jobs created.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Amazon SageMaker Runtime

The following actions are supported by Amazon SageMaker Runtime:

- InvokeEndpoint
- InvokeEndpointAsync
- InvokeEndpointWithResponseStream

InvokeEndpoint

Service: Amazon SageMaker Runtime

After you deploy a model into production using Amazon SageMaker AI hosting services, your client applications use this API to get inferences from the model hosted at the specified endpoint.

For an overview of Amazon SageMaker AI, see How It Works.

Amazon SageMaker AI strips all POST headers except those supported by the API. Amazon SageMaker AI might add additional headers. You should not rely on the behavior of headers outside those enumerated in the request syntax.

Calls to `InvokeEndpoint` are authenticated by using AWS Signature Version 4. For information, see Authenticating Requests (AWS Signature Version 4) in the *Amazon S3 API Reference*.

A customer's model containers must respond to requests within 60 seconds. The model itself can have a maximum processing time of 60 seconds before responding to invocations. If your model is going to take 50-60 seconds of processing time, the SDK socket timeout should be set to be 70 seconds.

> **Note**
>
> Endpoints are scoped to an individual account, and are not public. The URL does not contain the account ID, but Amazon SageMaker AI determines the account ID from the authentication token that is supplied by the caller.

Request Syntax

```
POST /endpoints/EndpointName/invocations HTTP/1.1
Content-Type: ContentType
Accept: Accept
X-Amzn-SageMaker-Custom-Attributes: CustomAttributes
X-Amzn-SageMaker-Target-Model: TargetModel
X-Amzn-SageMaker-Target-Variant: TargetVariant
X-Amzn-SageMaker-Target-Container-Hostname: TargetContainerHostname
X-Amzn-SageMaker-Inference-Id: InferenceId
X-Amzn-SageMaker-Enable-Explanations: EnableExplanations
X-Amzn-SageMaker-Inference-Component: InferenceComponentName
X-Amzn-SageMaker-Session-Id: SessionId
```

Body

URI Request Parameters

The request uses the following URI parameters.

Accept

The desired MIME type of the inference response from the model container.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

ContentType

The MIME type of the input data in the request body.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

CustomAttributes

Provides additional information about a request for an inference submitted to a model hosted at an Amazon SageMaker AI endpoint. The information is an opaque value that is forwarded verbatim. You could use this value, for example, to provide an ID that you can use to track a request or to provide other metadata that a service endpoint was programmed to process. The value must consist of no more than 1024 visible US-ASCII characters as specified in Section 3.3.6. Field Value Components of the Hypertext Transfer Protocol (HTTP/1.1).

The code in your model is responsible for setting or updating any custom attributes in the response. If your code does not set this value in the response, an empty value is returned. For example, if a custom attribute represents the trace ID, your model can prepend the custom attribute with `Trace ID:` in your post-processing function.

This feature is currently supported in the AWS SDKs but not in the Amazon SageMaker AI Python SDK.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

EnableExplanations

An optional JMESPath expression used to override the `EnableExplanations` parameter of the `ClarifyExplainerConfig` API. See the EnableExplanations section in the developer guide for more information.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.*`

EndpointName

The name of the endpoint that you specified when you created the endpoint using the CreateEndpoint API.

Length Constraints: Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

InferenceComponentName

If the endpoint hosts one or more inference components, this parameter specifies the name of inference component to invoke.

Length Constraints: Maximum length of 63.

Pattern: `^[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?$`

InferenceId

If you provide a value, it is added to the captured data when you enable data capture on the endpoint. For information about data capture, see Capture Data.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\A\S[\p{Print}]*\z`

SessionId

Creates a stateful session or identifies an existing one. You can do one of the following:

- Create a stateful session by specifying the value `NEW_SESSION`.
- Send your request to an existing stateful session by specifying the ID of that session.

With a stateful session, you can send multiple requests to a stateful model. When you create a session with a stateful model, the model must create the session ID and set the expiration time. The model must also provide that information in the response to your request. You can get the ID and timestamp from the `NewSessionId` response parameter. For any subsequent request where you specify that session ID, SageMaker AI routes the request to the same instance that supports the session.

Length Constraints: Maximum length of 256.

Pattern: `^(NEW_SESSION)$|^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

TargetContainerHostname

If the endpoint hosts multiple containers and is configured to use direct invocation, this parameter specifies the host name of the container to invoke.

Length Constraints: Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

TargetModel

The model to request for inference when invoking a multi-model endpoint.

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `\A\S[\p{Print}]*\z`

TargetVariant

Specify the production variant to send the inference request to when invoking an endpoint that is running two or more variants. Note that this parameter overrides the default behavior for the endpoint, which is to distribute the invocation traffic based on the variant weights.

For information about how to use variant targeting to perform a/b testing, see Test models in production

Length Constraints: Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Request Body

The request accepts the following binary data.

Body

Provides input data, in the format specified in the `ContentType` request header. Amazon SageMaker AI passes all of the data in the body to the model.

For information about the format of the request body, see Common Data Formats-Inference.

Length Constraints: Maximum length of 6291456.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-Type: ContentType
x-Amzn-Invoked-Production-Variant: InvokedProductionVariant
X-Amzn-SageMaker-Custom-Attributes: CustomAttributes
X-Amzn-SageMaker-New-Session-Id: NewSessionId
X-Amzn-SageMaker-Closed-Session-Id: ClosedSessionId

Body
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

ClosedSessionId

If you closed a stateful session with your request, the ID of that session.

Length Constraints: Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

ContentType

The MIME type of the inference returned from the model container.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

CustomAttributes

Provides additional information in the response about the inference returned by a model hosted at an Amazon SageMaker AI endpoint. The information is an opaque value that is forwarded verbatim. You could use this value, for example, to return an ID received in the `CustomAttributes` header of a request or other metadata that a service endpoint was programmed to produce. The value must consist of no more than 1024 visible US-ASCII characters as specified in Section 3.3.6. Field Value Components of the Hypertext Transfer Protocol (HTTP/1.1). If the customer wants the custom attribute returned, the model must set the custom attribute to be included on the way back.

The code in your model is responsible for setting or updating any custom attributes in the response. If your code does not set this value in the response, an empty value is returned. For example, if a custom attribute represents the trace ID, your model can prepend the custom attribute with `Trace ID:` in your post-processing function.

This feature is currently supported in the AWS SDKs but not in the Amazon SageMaker AI Python SDK.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

InvokedProductionVariant

Identifies the production variant that was invoked.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

NewSessionId

If you created a stateful session with your request, the ID and expiration time that the model assigns to that session.

Length Constraints: Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*;\sExpires=[0-9]{4}-[0-9]{2}-[0-9]{2}T[0-9]{2}:[0-9]{2}:[0-9]{2}Z$`

The response returns the following as the HTTP body.

Body

Includes the inference provided by the model.

For information about the format of the response body, see Common Data Formats-Inference.

If the explainer is activated, the body includes the explanations provided by the model. For more information, see the **Response section** under Invoke the Endpoint in the Developer Guide.

Length Constraints: Maximum length of 6291456.

Errors

For information about the errors that are common to all actions, see Common Error Types.

InternalDependencyException

Your request caused an exception with an internal dependency. Contact customer support.

HTTP Status Code: 530

InternalFailure

An internal failure occurred.

HTTP Status Code: 500

ModelError

Model (owned by the customer in the container) returned 4xx or 5xx error code.

LogStreamArn

The Amazon Resource Name (ARN) of the log stream.

OriginalMessage

Original message.

OriginalStatusCode

Original status code.

HTTP Status Code: 424

ModelNotReadyException

Either a serverless endpoint variant's resources are still being provisioned, or a multi-model endpoint is still downloading or loading the target model. Wait and try your request again.

HTTP Status Code: 429

ServiceUnavailable

The service is unavailable. Try your call again.

HTTP Status Code: 503

ValidationError

Inspect your request and try again.

HTTP Status Code: 400

Examples

Pass a trace ID in the CustomAttribute of a request and return it in the CustomAttribute of the response.

In this example a trace ID is passed to the service endpoint in the `CustomAttributes` header of the request and then retrieved and returned in the `CustomAttributes` header of the response.

Sample Request

```
import boto3

client = boto3.client('sagemaker-runtime')

custom_attributes = "c000b4f9-df62-4c85-a0bf-7c525f9104a4"  # An example of a trace ID.
endpoint_name = "..."                                       # Your endpoint name.
content_type = "..."                                        # The MIME type of the
 input data in the request body.
accept = "..."                                              # The desired MIME type of
 the inference in the response.
payload = "..."                                             # Payload for inference.
response = client.invoke_endpoint(
    EndpointName=endpoint_name,
    CustomAttributes=custom_attributes,
```

```
    ContentType=content_type,
    Accept=accept,
    Body=payload
    )

print(response['CustomAttributes'])                          # If model receives and
 updates the custom_attributes header
                                                             # by adding "Trace id: " in
 front of custom_attributes in the request,
                                                             # custom_attributes in
 response becomes
                                                             # "Trace ID: c000b4f9-
df62-4c85-a0bf-7c525f9104a4"
```

Sample Response

```
Trace ID: c000b4f9-df62-4c85-a0bf-7c525f9104a4
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

InvokeEndpointAsync

Service: Amazon SageMaker Runtime

After you deploy a model into production using Amazon SageMaker AI hosting services, your client applications use this API to get inferences from the model hosted at the specified endpoint in an asynchronous manner.

Inference requests sent to this API are enqueued for asynchronous processing. The processing of the inference request may or may not complete before you receive a response from this API. The response from this API will not contain the result of the inference request but contain information about where you can locate it.

Amazon SageMaker AI strips all POST headers except those supported by the API. Amazon SageMaker AI might add additional headers. You should not rely on the behavior of headers outside those enumerated in the request syntax.

Calls to `InvokeEndpointAsync` are authenticated by using AWS Signature Version 4. For information, see Authenticating Requests (AWS Signature Version 4) in the *Amazon S3 API Reference*.

Request Syntax

```
POST /endpoints/EndpointName/async-invocations HTTP/1.1
X-Amzn-SageMaker-Content-Type: ContentType
X-Amzn-SageMaker-Accept: Accept
X-Amzn-SageMaker-Custom-Attributes: CustomAttributes
X-Amzn-SageMaker-Inference-Id: InferenceId
X-Amzn-SageMaker-InputLocation: InputLocation
X-Amzn-SageMaker-S3OutputPathExtension: S3OutputPathExtension
X-Amzn-SageMaker-Filename: Filename
X-Amzn-SageMaker-RequestTTLSeconds: RequestTTLSeconds
X-Amzn-SageMaker-InvocationTimeoutSeconds: InvocationTimeoutSeconds

Body
```

URI Request Parameters

The request uses the following URI parameters.

Accept

The desired MIME type of the inference response from the model container.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

ContentType

The MIME type of the input data in the request body.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

CustomAttributes

Provides additional information about a request for an inference submitted to a model hosted at an Amazon SageMaker AI endpoint. The information is an opaque value that is forwarded verbatim. You could use this value, for example, to provide an ID that you can use to track a request or to provide other metadata that a service endpoint was programmed to process. The value must consist of no more than 1024 visible US-ASCII characters as specified in Section 3.3.6. Field Value Components of the Hypertext Transfer Protocol (HTTP/1.1).

The code in your model is responsible for setting or updating any custom attributes in the response. If your code does not set this value in the response, an empty value is returned. For example, if a custom attribute represents the trace ID, your model can prepend the custom attribute with `Trace ID:` in your post-processing function.

This feature is currently supported in the AWS SDKs but not in the Amazon SageMaker AI Python SDK.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

EndpointName

The name of the endpoint that you specified when you created the endpoint using the CreateEndpoint API.

Length Constraints: Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

Filename

The filename for the inference response payload stored in Amazon S3. If not specified, Amazon SageMaker AI generates a filename based on the inference ID.

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `^(?!.*\..*\.)[a-zA-Z0-9][a-zA-Z0-9-_\.]*$`

InferenceId

The identifier for the inference request. Amazon SageMaker AI will generate an identifier for you if none is specified.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\A\S[\p{Print}]*\z`

InputLocation

The Amazon S3 URI where the inference request payload is stored.

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `^(https|s3)://([^/]+)/?(.*)$`

InvocationTimeoutSeconds

Maximum amount of time in seconds a request can be processed before it is marked as expired. The default is 15 minutes, or 900 seconds.

Valid Range: Minimum value of 1. Maximum value of 3600.

RequestTTLSeconds

Maximum age in seconds a request can be in the queue before it is marked as expired. The default is 6 hours, or 21,600 seconds.

Valid Range: Minimum value of 60. Maximum value of 21600.

S3OutputPathExtension

The path extension that is appended to the Amazon S3 output path where the inference response payload is stored.

Length Constraints: Minimum length of 1. Maximum length of 512.

Pattern: `^(?!s3:|https:)[a-zA-Z0-9!_.*'()/-]+$`

Request Body

The request accepts the following binary data.

Body

Provides inline input data for the inference request, in the format specified in the `ContentType` request header. Use this parameter to send the request payload directly in the API call instead of uploading it to Amazon S3 and referencing it with `InputLocation`. The inline payload can be up to 128,000 bytes.

Body and `InputLocation` are mutually exclusive. Provide exactly one of them.

For information about the format of the request body, see Common Data Formats-Inference.

Length Constraints: Maximum length of 128000.

Response Syntax

```
HTTP/1.1 202
X-Amzn-SageMaker-OutputLocation: OutputLocation
X-Amzn-SageMaker-FailureLocation: FailureLocation
Content-type: application/json

{
   "InferenceId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The response returns the following HTTP headers.

FailureLocation

The Amazon S3 URI where the inference failure response payload is stored.

Length Constraints: Maximum length of 1024.

Pattern: \p{ASCII}*

OutputLocation

The Amazon S3 URI where the inference response payload is stored.

Length Constraints: Maximum length of 1024.

Pattern: \p{ASCII}*

The following data is returned in JSON format by the service.

InferenceId

Identifier for an inference request. This will be the same as the `InferenceId` specified in the input. Amazon SageMaker AI will generate an identifier for you if you do not specify one.

Type: String

Length Constraints: Maximum length of 1024.

Pattern: \p{ASCII}*

Errors

For information about the errors that are common to all actions, see Common Error Types.

InternalFailure

An internal failure occurred.

HTTP Status Code: 500

ServiceUnavailable

The service is unavailable. Try your call again.

HTTP Status Code: 503

ValidationError

Inspect your request and try again.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

InvokeEndpointWithResponseStream

Service: Amazon SageMaker Runtime

Invokes a model at the specified endpoint to return the inference response as a stream. The inference stream provides the response payload incrementally as a series of parts. Before you can get an inference stream, you must have access to a model that's deployed using Amazon SageMaker AI hosting services, and the container for that model must support inference streaming.

For more information that can help you use this API, see the following sections in the *Amazon SageMaker AI Developer Guide*:

- For information about how to add streaming support to a model, see How Containers Serve Requests.
- For information about how to process the streaming response, see Invoke real-time endpoints.

Before you can use this operation, your IAM permissions must allow the `sagemaker:InvokeEndpoint` action. For more information about Amazon SageMaker AI actions for IAM policies, see Actions, resources, and condition keys for Amazon SageMaker AI in the *IAM Service Authorization Reference*.

Amazon SageMaker AI strips all POST headers except those supported by the API. Amazon SageMaker AI might add additional headers. You should not rely on the behavior of headers outside those enumerated in the request syntax.

Calls to `InvokeEndpointWithResponseStream` are authenticated by using AWS Signature Version 4. For information, see Authenticating Requests (AWS Signature Version 4) in the *Amazon S3 API Reference*.

Request Syntax

```
POST /endpoints/EndpointName/invocations-response-stream HTTP/1.1
Content-Type: ContentType
X-Amzn-SageMaker-Accept: Accept
X-Amzn-SageMaker-Custom-Attributes: CustomAttributes
X-Amzn-SageMaker-Target-Variant: TargetVariant
X-Amzn-SageMaker-Target-Container-Hostname: TargetContainerHostname
X-Amzn-SageMaker-Inference-Id: InferenceId
X-Amzn-SageMaker-Inference-Component: InferenceComponentName
X-Amzn-SageMaker-Session-Id: SessionId
```

Body

URI Request Parameters

The request uses the following URI parameters.

Accept

The desired MIME type of the inference response from the model container.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

ContentType

The MIME type of the input data in the request body.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

CustomAttributes

Provides additional information about a request for an inference submitted to a model hosted at an Amazon SageMaker AI endpoint. The information is an opaque value that is forwarded verbatim. You could use this value, for example, to provide an ID that you can use to track a request or to provide other metadata that a service endpoint was programmed to process. The value must consist of no more than 1024 visible US-ASCII characters as specified in Section 3.3.6. Field Value Components of the Hypertext Transfer Protocol (HTTP/1.1).

The code in your model is responsible for setting or updating any custom attributes in the response. If your code does not set this value in the response, an empty value is returned. For example, if a custom attribute represents the trace ID, your model can prepend the custom attribute with `Trace ID:` in your post-processing function.

This feature is currently supported in the AWS SDKs but not in the Amazon SageMaker AI Python SDK.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

EndpointName

The name of the endpoint that you specified when you created the endpoint using the CreateEndpoint API.

Length Constraints: Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

InferenceComponentName

If the endpoint hosts one or more inference components, this parameter specifies the name of inference component to invoke for a streaming response.

Length Constraints: Maximum length of 63.

Pattern: `^[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?$`

InferenceId

An identifier that you assign to your request.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\A\S[\p{Print}]*\z`

SessionId

The ID of a stateful session to handle your request.

You can't create a stateful session by using the `InvokeEndpointWithResponseStream` action. Instead, you can create one by using the `InvokeEndpoint` action. In your request, you specify `NEW_SESSION` for the `SessionId` request parameter. The response to that request provides the session ID for the `NewSessionId` response parameter.

Length Constraints: Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

TargetContainerHostname

If the endpoint hosts multiple containers and is configured to use direct invocation, this parameter specifies the host name of the container to invoke.

Length Constraints: Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

TargetVariant

Specify the production variant to send the inference request to when invoking an endpoint that is running two or more variants. Note that this parameter overrides the default behavior for the endpoint, which is to distribute the invocation traffic based on the variant weights.

For information about how to use variant targeting to perform a/b testing, see Test models in production

Length Constraints: Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Request Body

The request accepts the following binary data.

Body

Provides input data, in the format specified in the `ContentType` request header. Amazon SageMaker AI passes all of the data in the body to the model.

For information about the format of the request body, see Common Data Formats-Inference.

Length Constraints: Maximum length of 6291456.

Required: Yes

Response Syntax

```
HTTP/1.1 200
X-Amzn-SageMaker-Content-Type: ContentType
x-Amzn-Invoked-Production-Variant: InvokedProductionVariant
X-Amzn-SageMaker-Custom-Attributes: CustomAttributes
Content-type: application/json

{
   "InternalStreamFailure": {
```

```
    },
    "ModelStreamError": {
    },
    "PayloadPart": {
       "Bytes": blob
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

ContentType

The MIME type of the inference returned from the model container.

Length Constraints: Maximum length of 1024.

Pattern: `\p{ASCII}*`

CustomAttributes

Provides additional information in the response about the inference returned by a model hosted at an Amazon SageMaker AI endpoint. The information is an opaque value that is forwarded verbatim. You could use this value, for example, to return an ID received in the `CustomAttributes` header of a request or other metadata that a service endpoint was programmed to produce. The value must consist of no more than 1024 visible US-ASCII characters as specified in Section 3.3.6. Field Value Components of the Hypertext Transfer Protocol (HTTP/1.1). If the customer wants the custom attribute returned, the model must set the custom attribute to be included on the way back.

The code in your model is responsible for setting or updating any custom attributes in the response. If your code does not set this value in the response, an empty value is returned. For example, if a custom attribute represents the trace ID, your model can prepend the custom attribute with `Trace ID:` in your post-processing function.

This feature is currently supported in the AWS SDKs but not in the Amazon SageMaker AI Python SDK.

Length Constraints: Maximum length of 1024.

Pattern: \p{ASCII}*

InvokedProductionVariant

Identifies the production variant that was invoked.

Length Constraints: Maximum length of 1024.

Pattern: \p{ASCII}*

The following data is returned in JSON format by the service.

InternalStreamFailure

The stream processing failed because of an unknown error, exception or failure. Try your request again.

Type: Exception

HTTP Status Code:

ModelStreamError

An error occurred while streaming the response body. This error can have the following error codes:

ModelInvocationTimeExceeded

The model failed to finish sending the response within the timeout period allowed by Amazon SageMaker AI.

StreamBroken

The Transmission Control Protocol (TCP) connection between the client and the model was reset or closed.

Type: Exception

HTTP Status Code:

PayloadPart

A wrapper for pieces of the payload that's returned in response to a streaming inference request. A streaming inference response consists of one or more payload parts.

Type: PayloadPart object

Errors

For information about the errors that are common to all actions, see Common Error Types.

InternalFailure

An internal failure occurred.

HTTP Status Code: 500

InternalStreamFailure

The stream processing failed because of an unknown error, exception or failure. Try your request again.

HTTP Status Code: 500

ModelError

Model (owned by the customer in the container) returned 4xx or 5xx error code.

LogStreamArn

The Amazon Resource Name (ARN) of the log stream.

OriginalMessage

Original message.

OriginalStatusCode

Original status code.

HTTP Status Code: 424

ModelStreamError

An error occurred while streaming the response body. This error can have the following error codes:

ModelInvocationTimeExceeded

The model failed to finish sending the response within the timeout period allowed by Amazon SageMaker AI.

StreamBroken

The Transmission Control Protocol (TCP) connection between the client and the model was reset or closed.

ErrorCode

This error can have the following error codes:

ModelInvocationTimeExceeded

The model failed to finish sending the response within the timeout period allowed by Amazon SageMaker AI.

StreamBroken

The Transmission Control Protocol (TCP) connection between the client and the model was reset or closed.

HTTP Status Code: 400

ServiceUnavailable

The service is unavailable. Try your call again.

HTTP Status Code: 503

ValidationError

Inspect your request and try again.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Amazon Sagemaker Edge Manager

The following actions are supported by Amazon Sagemaker Edge Manager:

- GetDeployments
- GetDeviceRegistration
- SendHeartbeat

GetDeployments

Service: Amazon Sagemaker Edge Manager

Use to get the active deployments from a device.

Request Syntax

```
POST /GetDeployments HTTP/1.1
Content-type: application/json

{
   "DeviceFleetName": "string",
   "DeviceName": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

DeviceFleetName

The name of the fleet to which the device belongs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*_*[a-zA-Z0-9])*$`

Required: Yes

DeviceName

The unique name of the device from which you want to get the configuration of active deployments.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*_*[a-zA-Z0-9])*$`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "Deployments": [
      {
         "Definitions": [
            {
               "Checksum": {
                  "Sum": "string",
                  "Type": "string"
               },
               "ModelHandle": "string",
               "S3Url": "string",
               "State": "string"
            }
         ],
         "DeploymentName": "string",
         "FailureHandlingPolicy": "string",
         "Type": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Deployments

Returns a list of the configurations of the active deployments on the device.

Type: Array of EdgeDeployment objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

InternalServiceException

An internal failure occurred. Try your request again. If the problem persists, contact AWS customer support.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetDeviceRegistration

Service: Amazon Sagemaker Edge Manager

Use to check if a device is registered with SageMaker Edge Manager.

Request Syntax

```
POST /GetDeviceRegistration HTTP/1.1
Content-type: application/json

{
   "DeviceFleetName": "string",
   "DeviceName": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

DeviceFleetName

The name of the fleet to which the device belongs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*_*[a-zA-Z0-9])*$`

Required: Yes

DeviceName

The unique name of the device from which you want to get the registration status.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*_*[a-zA-Z0-9])*$`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "CacheTTL": "string",
   "DeviceRegistration": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

CacheTTL

The amount of time, in seconds, that the registration status is stored on the device's cache before it is refreshed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1000.

DeviceRegistration

Describes if the device is currently registered with SageMaker Edge Manager.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1000.

Errors

For information about the errors that are common to all actions, see Common Error Types.

InternalServiceException

An internal failure occurred. Try your request again. If the problem persists, contact AWS customer support.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

SendHeartbeat

Service: Amazon Sagemaker Edge Manager

Use to get the current status of devices registered on SageMaker Edge Manager.

Request Syntax

```
POST /SendHeartbeat HTTP/1.1
Content-type: application/json

{
   "AgentMetrics": [
      {
         "Dimension": "string",
         "MetricName": "string",
         "Timestamp": number,
         "Value": number
      }
   ],
   "AgentVersion": "string",
   "DeploymentResult": {
      "DeploymentEndTime": number,
      "DeploymentModels": [
         {
            "DesiredState": "string",
            "ModelHandle": "string",
            "ModelName": "string",
            "ModelVersion": "string",
            "RollbackFailureReason": "string",
            "State": "string",
            "Status": "string",
            "StatusReason": "string"
         }
      ],
      "DeploymentName": "string",
      "DeploymentStartTime": number,
      "DeploymentStatus": "string",
      "DeploymentStatusMessage": "string"
   },
   "DeviceFleetName": "string",
   "DeviceName": "string",
   "Models": [
      {
```

```
         "LatestInference": number,
         "LatestSampleTime": number,
         "ModelMetrics": [
            {
               "Dimension": "string",
               "MetricName": "string",
               "Timestamp": number,
               "Value": number
            }
         ],
         "ModelName": "string",
         "ModelVersion": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

AgentMetrics

For internal use. Returns a list of SageMaker Edge Manager agent operating metrics.

Type: Array of EdgeMetric objects

Required: No

AgentVersion

Returns the version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9\ \_\.]+`

Required: Yes

DeploymentResult

Returns the result of a deployment on the device.

Type: DeploymentResult object

Required: No

DeviceFleetName

The name of the fleet to which the device belongs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*_*[a-zA-Z0-9])*$`

Required: Yes

DeviceName

The unique name of the device.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*_*[a-zA-Z0-9])*$`

Required: Yes

Models

Returns a list of models deployed on the the device.

Type: Array of Model objects

Required: No

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

InternalServiceException

An internal failure occurred. Try your request again. If the problem persists, contact AWS customer support.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Amazon SageMaker Feature Store Runtime

The following actions are supported by Amazon SageMaker Feature Store Runtime:

- BatchGetRecord

- DeleteRecord
- GetRecord
- PutRecord

BatchGetRecord

Service: Amazon SageMaker Feature Store Runtime

Retrieves a batch of `Records` from a `FeatureGroup`.

Request Syntax

```
POST /BatchGetRecord HTTP/1.1
Content-type: application/json

{
   "ExpirationTimeResponse": "string",
   "Identifiers": [
      {
         "FeatureGroupName": "string",
         "FeatureNames": [ "string" ],
         "RecordIdentifiersValueAsString": [ "string" ]
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

ExpirationTimeResponse

Parameter to request `ExpiresAt` in response. If `Enabled`, `BatchGetRecord` will return the value of `ExpiresAt`, if it is not null. If `Disabled` and null, `BatchGetRecord` will return null.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

Identifiers

A list containing the name or Amazon Resource Name (ARN) of the `FeatureGroup`, the list of names of `Features` to be retrieved, and the corresponding `RecordIdentifier` values as strings.

Type: Array of BatchGetRecordIdentifier objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "Errors": [
      {
         "ErrorCode": "string",
         "ErrorMessage": "string",
         "FeatureGroupName": "string",
         "RecordIdentifierValueAsString": "string"
      }
   ],
   "Records": [
      {
         "ExpiresAt": "string",
         "FeatureGroupName": "string",
         "Record": [
            {
               "FeatureName": "string",
               "ValueAsString": "string",
               "ValueAsStringList": [ "string" ]
            }
         ],
         "RecordIdentifierValueAsString": "string"
      }
   ],
   "UnprocessedIdentifiers": [
      {
         "FeatureGroupName": "string",
```

```
            "FeatureNames": [ "string" ],
            "RecordIdentifiersValueAsString": [ "string" ]
        }
    ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Errors

A list of errors that have occurred when retrieving a batch of Records.

Type: Array of BatchGetRecordError objects

Array Members: Minimum number of 0 items.

Records

A list of Records you requested to be retrieved in batch.

Type: Array of BatchGetRecordResultDetail objects

Array Members: Minimum number of 0 items.

UnprocessedIdentifiers

A unprocessed list of `FeatureGroup` names, with their corresponding `RecordIdentifier` value, and Feature name.

Type: Array of BatchGetRecordIdentifier objects

Array Members: Minimum number of 0 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessForbidden

You do not have permission to perform an action.

HTTP Status Code: 403

InternalFailure

An internal failure occurred. Try your request again. If the problem persists, contact AWS customer support.

HTTP Status Code: 500

ServiceUnavailable

The service is currently unavailable.

HTTP Status Code: 503

ValidationError

There was an error validating your request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteRecord

Service: Amazon SageMaker Feature Store Runtime

Deletes a `Record` from a `FeatureGroup` in the `OnlineStore`. Feature Store supports both `SoftDelete` and `HardDelete`. For `SoftDelete` (default), feature columns are set to `null` and the record is no longer retrievable by `GetRecord` or `BatchGetRecord`. For `HardDelete`, the complete `Record` is removed from the `OnlineStore`. In both cases, Feature Store appends the deleted record marker to the `OfflineStore`. The deleted record marker is a record with the same `RecordIdentifer` as the original, but with `is_deleted` value set to `True`, `EventTime` set to the delete input `EventTime`, and other feature values set to `null`.

Note that the `EventTime` specified in `DeleteRecord` should be set later than the `EventTime` of the existing record in the `OnlineStore` for that `RecordIdentifer`. If it is not, the deletion does not occur:

- For `SoftDelete`, the existing (not deleted) record remains in the `OnlineStore`, though the delete record marker is still written to the `OfflineStore`.
- `HardDelete` returns `EventTime: 400 ValidationException` to indicate that the delete operation failed. No delete record marker is written to the `OfflineStore`.

When a record is deleted from the `OnlineStore`, the deleted record marker is appended to the `OfflineStore`. If you have the Iceberg table format enabled for your `OfflineStore`, you can remove all history of a record from the `OfflineStore` using Amazon Athena or Apache Spark. For information on how to hard delete a record from the `OfflineStore` with the Iceberg table format enabled, see Delete records from the offline store.

Request Syntax

```
DELETE /FeatureGroup/FeatureGroupName?
DeletionMode=DeletionMode&EventTime=EventTime&RecordIdentifierValueAsString=RecordIdentifierVal
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

DeletionMode

The name of the deletion mode for deleting the record. By default, the deletion mode is set to `SoftDelete`.

Valid Values: `SoftDelete | HardDelete`

EventTime

Timestamp indicating when the deletion event occurred. `EventTime` can be used to query data at a certain point in time.

Length Constraints: Maximum length of 358400.

Pattern: `.*`

Required: Yes

FeatureGroupName

The name or Amazon Resource Name (ARN) of the feature group to delete the record from.

Length Constraints: Minimum length of 1. Maximum length of 150.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group/)?([a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63})`

Required: Yes

RecordIdentifierValueAsString

The value for the `RecordIdentifier` that uniquely identifies the record, in string format.

Length Constraints: Maximum length of 358400.

Pattern: `.*`

Required: Yes

TargetStores

A list of stores from which you're deleting the record. By default, Feature Store deletes the record from all of the stores that you're using for the `FeatureGroup`.

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `OnlineStore | OfflineStore`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessForbidden

You do not have permission to perform an action.

HTTP Status Code: 403

InternalFailure

An internal failure occurred. Try your request again. If the problem persists, contact AWS customer support.

HTTP Status Code: 500

ServiceUnavailable

The service is currently unavailable.

HTTP Status Code: 503

ValidationError

There was an error validating your request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetRecord

Service: Amazon SageMaker Feature Store Runtime

Use for `OnlineStore` serving from a `FeatureStore`. Only the latest records stored in the `OnlineStore` can be retrieved. If no Record with `RecordIdentifierValue` is found, then an empty result is returned.

Request Syntax

```
GET /FeatureGroup/FeatureGroupName?
ExpirationTimeResponse=ExpirationTimeResponse&FeatureName=FeatureNames&RecordIdentifierValueAsS
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

ExpirationTimeResponse

Parameter to request `ExpiresAt` in response. If `Enabled`, `GetRecord` will return the value of `ExpiresAt`, if it is not null. If `Disabled` and null, `GetRecord` will return null.

Valid Values: `Enabled | Disabled`

FeatureGroupName

The name or Amazon Resource Name (ARN) of the feature group from which you want to retrieve a record.

Length Constraints: Minimum length of 1. Maximum length of 150.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group/)?([a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63})`

Required: Yes

FeatureNames

List of names of Features to be retrieved. If not specified, the latest value for all the Features are returned.

Array Members: Minimum number of 1 item.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `^[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

RecordIdentifierValueAsString

The value that corresponds to `RecordIdentifier` type and uniquely identifies the record in the `FeatureGroup`.

Length Constraints: Maximum length of 358400.

Pattern: `.*`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "ExpiresAt": "string",
   "Record": [
      {
         "FeatureName": "string",
         "ValueAsString": "string",
         "ValueAsStringList": [ "string" ]
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ExpiresAt

The `ExpiresAt` ISO string of the requested record.

Type: String

Record

The record you requested. A list of `FeatureValues`.

Type: Array of FeatureValue objects

Array Members: Minimum number of 1 item.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessForbidden

You do not have permission to perform an action.

HTTP Status Code: 403

InternalFailure

An internal failure occurred. Try your request again. If the problem persists, contact AWS customer support.

HTTP Status Code: 500

ResourceNotFound

A resource that is required to perform an action was not found.

HTTP Status Code: 404

ServiceUnavailable

The service is currently unavailable.

HTTP Status Code: 503

ValidationError

There was an error validating your request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

PutRecord

Service: Amazon SageMaker Feature Store Runtime

The `PutRecord` API is used to ingest a list of `Records` into your feature group.

If a new record's `EventTime` is greater, the new record is written to both the `OnlineStore` and `OfflineStore`. Otherwise, the record is a historic record and it is written only to the `OfflineStore`.

You can specify the ingestion to be applied to the `OnlineStore`, `OfflineStore`, or both by using the `TargetStores` request parameter.

You can set the ingested record to expire at a given time to live (TTL) duration after the record's event time, `ExpiresAt` = `EventTime` + `TtlDuration`, by specifying the `TtlDuration` parameter. A record level `TtlDuration` is set when specifying the `TtlDuration` parameter using the `PutRecord` API call. If the input `TtlDuration` is `null` or unspecified, `TtlDuration` is set to the default feature group level `TtlDuration`. A record level `TtlDuration` supersedes the group level `TtlDuration`.

Request Syntax

```
PUT /FeatureGroup/FeatureGroupName HTTP/1.1
Content-type: application/json

{
   "Record": [
      {
         "FeatureName": "string",
         "ValueAsString": "string",
         "ValueAsStringList": [ "string" ]
      }
   ],
   "TargetStores": [ "string" ],
   "TtlDuration": {
      "Unit": "string",
      "Value": number
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

FeatureGroupName

The name or Amazon Resource Name (ARN) of the feature group that you want to insert the record into.

Length Constraints: Minimum length of 1. Maximum length of 150.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group/)?([a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

Record

List of FeatureValues to be inserted. This will be a full over-write. If you only want to update few of the feature values, do the following:

- Use `GetRecord` to retrieve the latest record.
- Update the record returned from `GetRecord`.
- Use `PutRecord` to update feature values.

Type: Array of FeatureValue objects

Array Members: Minimum number of 1 item.

Required: Yes

TargetStores

A list of stores to which you're adding the record. By default, Feature Store adds the record to all of the stores that you're using for the `FeatureGroup`.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `OnlineStore | OfflineStore`

Required: No

TtlDuration

Time to live duration, where the record is hard deleted after the expiration time is reached; `ExpiresAt` = `EventTime` + `TtlDuration`. For information on HardDelete, see the DeleteRecord API in the Amazon SageMaker API Reference guide.

Type: TtlDuration object

Required: No

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessForbidden

You do not have permission to perform an action.

HTTP Status Code: 403

InternalFailure

An internal failure occurred. Try your request again. If the problem persists, contact AWS customer support.

HTTP Status Code: 500

ServiceUnavailable

The service is currently unavailable.

HTTP Status Code: 503

ValidationError

There was an error validating your request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Amazon SageMaker geospatial capabilities

The following actions are supported by Amazon SageMaker geospatial capabilities:

- DeleteEarthObservationJob
- DeleteVectorEnrichmentJob
- ExportEarthObservationJob
- ExportVectorEnrichmentJob
- GetEarthObservationJob
- GetRasterDataCollection
- GetTile

- GetVectorEnrichmentJob
- ListEarthObservationJobs
- ListRasterDataCollections
- ListTagsForResource
- ListVectorEnrichmentJobs
- SearchRasterDataCollection
- StartEarthObservationJob
- StartVectorEnrichmentJob
- StopEarthObservationJob
- StopVectorEnrichmentJob
- TagResource
- UntagResource

DeleteEarthObservationJob

Service: Amazon SageMaker geospatial capabilities

Use this operation to delete an Earth Observation job.

Request Syntax

```
DELETE /earth-observation-jobs/Arn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

Arn

The Amazon Resource Name (ARN) of the Earth Observation job being deleted.

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:earth-observation-job/[a-z0-9]{12,}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

ConflictException

Updating or deleting a resource can cause an inconsistent state.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 409

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteVectorEnrichmentJob

Service: Amazon SageMaker geospatial capabilities

Use this operation to delete a Vector Enrichment job.

Request Syntax

```
DELETE /vector-enrichment-jobs/Arn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

Arn

The Amazon Resource Name (ARN) of the Vector Enrichment job being deleted.

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:vector-enrichment-job/[a-z0-9]{12,}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

ConflictException

Updating or deleting a resource can cause an inconsistent state.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 409

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ExportEarthObservationJob

Service: Amazon SageMaker geospatial capabilities

Use this operation to export results of an Earth Observation job and optionally source images used as input to the EOJ to an Amazon S3 location.

Request Syntax

```
POST /export-earth-observation-job HTTP/1.1
Content-type: application/json

{
   "Arn": "string",
   "ClientToken": "string",
   "ExecutionRoleArn": "string",
   "ExportSourceImages": boolean,
   "OutputConfig": {
      "S3Data": {
         "KmsKeyId": "string",
         "S3Uri": "string"
      }
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

Arn

The input Amazon Resource Name (ARN) of the Earth Observation job being exported.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:earth-observation-job/[a-z0-9]{12,}`

Required: Yes

ClientToken

A unique token that guarantees that the call to this API is idempotent.

Type: String

Length Constraints: Minimum length of 36. Maximum length of 64.

Required: No

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you specified for the job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:(aws[a-z-]*):iam::([0-9]{12}):role/[a-zA-Z0-9+=,.@_/-]+`

Required: Yes

ExportSourceImages

The source images provided to the Earth Observation job being exported.

Type: Boolean

Required: No

OutputConfig

An object containing information about the output file.

Type: OutputConfigInput object

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "Arn": "string",
```

```
   "CreationTime": "string",
   "ExecutionRoleArn": "string",
   "ExportSourceImages": boolean,
   "ExportStatus": "string",
   "OutputConfig": {
      "S3Data": {
         "KmsKeyId": "string",
         "S3Uri": "string"
      }
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The output Amazon Resource Name (ARN) of the Earth Observation job being exported.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:earth-observation-job/[a-z0-9]{12,}`

CreationTime

The creation time.

Type: Timestamp

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you specified for the job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:(aws[a-z-]*):iam::([0-9]{12}):role/[a-zA-Z0-9+=,.@_/-]+`

ExportSourceImages

The source images provided to the Earth Observation job being exported.

Type: Boolean

ExportStatus

The status of the results of the Earth Observation job being exported.

Type: String

Valid Values: `IN_PROGRESS | SUCCEEDED | FAILED`

OutputConfig

An object containing information about the output file.

Type: OutputConfigInput object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

ConflictException

Updating or deleting a resource can cause an inconsistent state.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 409

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ServiceQuotaExceededException

You have exceeded the service quota.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 402

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ExportVectorEnrichmentJob

Service: Amazon SageMaker geospatial capabilities

Use this operation to copy results of a Vector Enrichment job to an Amazon S3 location.

Request Syntax

```
POST /export-vector-enrichment-jobs HTTP/1.1
Content-type: application/json

{
   "Arn": "string",
   "ClientToken": "string",
   "ExecutionRoleArn": "string",
   "OutputConfig": {
      "S3Data": {
         "KmsKeyId": "string",
         "S3Uri": "string"
      }
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

Arn

The Amazon Resource Name (ARN) of the Vector Enrichment job.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:vector-enrichment-job/[a-z0-9]{12,}`

Required: Yes

ClientToken

A unique token that guarantees that the call to this API is idempotent.

Type: String

Length Constraints: Minimum length of 36. Maximum length of 64.

Required: No

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM rolewith permission to upload to the location in OutputConfig.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:(aws[a-z-]*):iam::([0-9]{12}):role/[a-zA-Z0-9+=,.@_/-]+`

Required: Yes

OutputConfig

Output location information for exporting Vector Enrichment Job results.

Type: ExportVectorEnrichmentJobOutputConfig object

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "Arn": "string",
   "CreationTime": "string",
   "ExecutionRoleArn": "string",
   "ExportStatus": "string",
   "OutputConfig": {
      "S3Data": {
         "KmsKeyId": "string",
         "S3Uri": "string"
      }
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The Amazon Resource Name (ARN) of the Vector Enrichment job being exported.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:vector-enrichment-job/[a-z0-9]{12,}`

CreationTime

The creation time.

Type: Timestamp

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role with permission to upload to the location in OutputConfig.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:(aws[a-z-]*):iam::([0-9]{12}):role/[a-zA-Z0-9+=,.@_/-]+`

ExportStatus

The status of the results the Vector Enrichment job being exported.

Type: String

Valid Values: `IN_PROGRESS | SUCCEEDED | FAILED`

OutputConfig

Output location information for exporting Vector Enrichment Job results.

Type: ExportVectorEnrichmentJobOutputConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

ConflictException

Updating or deleting a resource can cause an inconsistent state.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 409

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ServiceQuotaExceededException

You have exceeded the service quota.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 402

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetEarthObservationJob

Service: Amazon SageMaker geospatial capabilities

Get the details for a previously initiated Earth Observation job.

Request Syntax

```
GET /earth-observation-jobs/Arn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

Arn

The Amazon Resource Name (ARN) of the Earth Observation job.

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:earth-observation-job/[a-z0-9]{12,}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "Arn": "string",
   "CreationTime": "string",
   "DurationInSeconds": number,
   "ErrorDetails": {
      "Message": "string",
      "Type": "string"
   },
   "ExecutionRoleArn": "string",
   "ExportErrorDetails": {
      "ExportResults": {
```

```
            "Message": "string",
            "Type": "string"
         },
         "ExportSourceImages": {
            "Message": "string",
            "Type": "string"
         }
      },
      "ExportStatus": "string",
      "InputConfig": {
         "PreviousEarthObservationJobArn": "string",
         "RasterDataCollectionQuery": {
            "AreaOfInterest": { ... },
            "PropertyFilters": {
               "LogicalOperator": "string",
               "Properties": [
                  {
                     "Property": { ... }
                  }
               ]
            },
            "RasterDataCollectionArn": "string",
            "RasterDataCollectionName": "string",
            "TimeRangeFilter": {
               "EndTime": "string",
               "StartTime": "string"
            }
         }
      },
      "JobConfig": { ... },
      "KmsKeyId": "string",
      "Name": "string",
      "OutputBands": [
         {
            "BandName": "string",
            "OutputDataType": "string"
         }
      ],
      "Status": "string",
      "Tags": {
         "string" : "string"
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The Amazon Resource Name (ARN) of the Earth Observation job.

Type: String

CreationTime

The creation time of the initiated Earth Observation job.

Type: Timestamp

DurationInSeconds

The duration of Earth Observation job, in seconds.

Type: Integer

ErrorDetails

Details about the errors generated during the Earth Observation job.

Type: EarthObservationJobErrorDetails object

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you specified for the job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:(aws[a-z-]*):iam::([0-9]{12}):role/[a-zA-Z0-9+=,.@_/-]+`

ExportErrorDetails

Details about the errors generated during ExportEarthObservationJob.

Type: ExportErrorDetails object

ExportStatus

The status of the Earth Observation job.

Type: String

Valid Values: `IN_PROGRESS | SUCCEEDED | FAILED`

InputConfig

Input data for the Earth Observation job.

Type: InputConfigOutput object

JobConfig

An object containing information about the job configuration.

Type: JobConfigInput object

Note: This object is a Union. Only one member of this object can be specified or returned.

KmsKeyId

The Key Management Service key ID for server-side encryption.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Name

The name of the Earth Observation job.

Type: String

OutputBands

Bands available in the output of an operation.

Type: Array of OutputBand objects

Status

The status of a previously initiated Earth Observation job.

Type: String

Valid Values: `INITIALIZING | IN_PROGRESS | STOPPING | COMPLETED | STOPPED | FAILED | DELETING | DELETED`

Tags

Each tag consists of a key and a value.

Type: String to string map

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetRasterDataCollection

Service: Amazon SageMaker geospatial capabilities

Use this operation to get details of a specific raster data collection.

Request Syntax

```
GET /raster-data-collection/Arn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

Arn

The Amazon Resource Name (ARN) of the raster data collection.

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:raster-data-collection/(public|premium|user)/[a-z0-9]{12,}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "Arn": "string",
   "Description": "string",
   "DescriptionPageUrl": "string",
   "ImageSourceBands": [ "string" ],
   "Name": "string",
   "SupportedFilters": [
      {
         "Maximum": number,
         "Minimum": number,
```

```
            "Name": "string",
            "Type": "string"
        }
    ],
    "Tags": {
        "string" : "string"
    },
    "Type": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The Amazon Resource Name (ARN) of the raster data collection.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:raster-data-collection/(public|premium|user)/[a-z0-9]{12,}`

Description

A description of the raster data collection.

Type: String

DescriptionPageUrl

The URL of the description page.

Type: String

ImageSourceBands

The list of image source bands in the raster data collection.

Type: Array of strings

Name

The name of the raster data collection.

Type: String

SupportedFilters

The filters supported by the raster data collection.

Type: Array of Filter objects

Tags

Each tag consists of a key and a value.

Type: String to string map

Type

The raster data collection type.

Type: String

Valid Values: `PUBLIC | PREMIUM | USER`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetTile

Service: Amazon SageMaker geospatial capabilities

Gets a web mercator tile for the given Earth Observation job.

Request Syntax

```
GET /tile/z/x/y?
Arn=Arn&ExecutionRoleArn=ExecutionRoleArn&ImageAssets=ImageAssets&ImageMask=ImageMask&OutputDat
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

Arn

The Amazon Resource Name (ARN) of the tile operation.

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:earth-observation-job/[a-z0-9]{12,}`

Required: Yes

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you specify.

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:(aws[a-z-]*):iam::([0-9]{12}):role/[a-zA-Z0-9+=,.@_/-]+`

ImageAssets

The particular assets or bands to tile.

Array Members: Minimum number of 1 item.

Required: Yes

ImageMask

Determines whether or not to return a valid data mask.

OutputDataType

The output data type of the tile operation.

Valid Values: `INT32 | FLOAT32 | INT16 | FLOAT64 | UINT16`

OutputFormat

The data format of the output tile. The formats include .npy, .png and .jpg.

PropertyFilters

Property filters for the imagery to tile.

Target

Determines what part of the Earth Observation job to tile. 'INPUT' or 'OUTPUT' are the valid options.

Valid Values: `INPUT | OUTPUT`

Required: Yes

TimeRangeFilter

Time range filter applied to imagery to find the images to tile.

x

The x coordinate of the tile input.

Required: Yes

y

The y coordinate of the tile input.

Required: Yes

z

The z coordinate of the tile input.

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200

BinaryFile
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following as the HTTP body.

BinaryFile

The output binary file.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetVectorEnrichmentJob

Service: Amazon SageMaker geospatial capabilities

Retrieves details of a Vector Enrichment Job for a given job Amazon Resource Name (ARN).

Request Syntax

```
GET /vector-enrichment-jobs/Arn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

Arn

The Amazon Resource Name (ARN) of the Vector Enrichment job.

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:vector-enrichment-job/[a-z0-9]{12,}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "Arn": "string",
   "CreationTime": "string",
   "DurationInSeconds": number,
   "ErrorDetails": {
      "ErrorMessage": "string",
      "ErrorType": "string"
   },
   "ExecutionRoleArn": "string",
   "ExportErrorDetails": {
```

```
      "Message": "string",
      "Type": "string"
   },
   "ExportStatus": "string",
   "InputConfig": {
      "DataSourceConfig": { ... },
      "DocumentType": "string"
   },
   "JobConfig": { ... },
   "KmsKeyId": "string",
   "Name": "string",
   "Status": "string",
   "Tags": {
      "string" : "string"
   },
   "Type": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The Amazon Resource Name (ARN) of the Vector Enrichment job.

Type: String

CreationTime

The creation time.

Type: Timestamp

DurationInSeconds

The duration of the Vector Enrichment job, in seconds.

Type: Integer

ErrorDetails

Details about the errors generated during the Vector Enrichment job.

Type: VectorEnrichmentJobErrorDetails object

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you specified for the job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:(aws[a-z-]*):iam::([0-9]{12}):role/[a-zA-Z0-9+=,.@_/-]+`

ExportErrorDetails

Details about the errors generated during the ExportVectorEnrichmentJob.

Type: VectorEnrichmentJobExportErrorDetails object

ExportStatus

The export status of the Vector Enrichment job being initiated.

Type: String

Valid Values: `IN_PROGRESS | SUCCEEDED | FAILED`

InputConfig

Input configuration information for the Vector Enrichment job.

Type: VectorEnrichmentJobInputConfig object

JobConfig

An object containing information about the job configuration.

Type: VectorEnrichmentJobConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

KmsKeyId

The Key Management Service key ID for server-side encryption.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Name

The name of the Vector Enrichment job.

Type: String

Status

The status of the initiated Vector Enrichment job.

Type: String

Valid Values: `INITIALIZING | IN_PROGRESS | STOPPING | STOPPED | COMPLETED | FAILED | DELETING | DELETED`

Tags

Each tag consists of a key and a value.

Type: String to string map

Type

The type of the Vector Enrichment job being initiated.

Type: String

Valid Values: `REVERSE_GEOCODING | MAP_MATCHING`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

ListEarthObservationJobs

Service: Amazon SageMaker geospatial capabilities

Use this operation to get a list of the Earth Observation jobs associated with the calling AWS account.

Request Syntax

```
POST /list-earth-observation-jobs HTTP/1.1
Content-type: application/json

{
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

MaxResults

The total number of items to return.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 20.

Required: No

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Required: No

SortBy

The parameter by which to sort the results.

Type: String

Required: No

SortOrder

An optional value that specifies whether you want the results sorted in `Ascending` or `Descending` order.

Type: String

Valid Values: `ASCENDING | DESCENDING`

Required: No

StatusEquals

A filter that retrieves only jobs with a specific status.

Type: String

Valid Values: `INITIALIZING | IN_PROGRESS | STOPPING | COMPLETED | STOPPED | FAILED | DELETING | DELETED`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "EarthObservationJobSummaries": [
      {
         "Arn": "string",
```

```
         "CreationTime": "string",
         "DurationInSeconds": number,
         "Name": "string",
         "OperationType": "string",
         "Status": "string",
         "Tags": {
            "string" : "string"
         }
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

EarthObservationJobSummaries

Contains summary information about the Earth Observation jobs.

Type: Array of ListEarthObservationJobOutputConfig objects

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListRasterDataCollections

Service: Amazon SageMaker geospatial capabilities

Use this operation to get raster data collections.

Request Syntax

```
GET /raster-data-collections?MaxResults=MaxResults&NextToken=NextToken HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

MaxResults

The total number of items to return.

Valid Range: Minimum value of 1. Maximum value of 20.

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Length Constraints: Minimum length of 0. Maximum length of 8192.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "NextToken": "string",
   "RasterDataCollectionSummaries": [
      {
         "Arn": "string",
         "Description": "string",
```

```
         "DescriptionPageUrl": "string",
         "Name": "string",
         "SupportedFilters": [
            {
               "Maximum": number,
               "Minimum": number,
               "Name": "string",
               "Type": "string"
            }
         ],
         "Tags": {
            "string" : "string"
         },
         "Type": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

RasterDataCollectionSummaries

Contains summary information about the raster data collection.

Type: Array of RasterDataCollectionMetadata objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListTagsForResource

Service: Amazon SageMaker geospatial capabilities

Lists the tags attached to the resource.

Request Syntax

```
GET /tags/ResourceArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

ResourceArn

The Amazon Resource Name (ARN) of the resource you want to tag.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "Tags": {
      "string" : "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Tags

Each tag consists of a key and a value.

Type: String to string map

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListVectorEnrichmentJobs

Service: Amazon SageMaker geospatial capabilities

Retrieves a list of vector enrichment jobs.

Request Syntax

```
POST /list-vector-enrichment-jobs HTTP/1.1
Content-type: application/json

{
   "MaxResults": number,
   "NextToken": "string",
   "SortBy": "string",
   "SortOrder": "string",
   "StatusEquals": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

MaxResults

The maximum number of items to return.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 20.

Required: No

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Required: No

SortBy

The parameter by which to sort the results.

Type: String

Required: No

SortOrder

An optional value that specifies whether you want the results sorted in `Ascending` or `Descending` order.

Type: String

Valid Values: `ASCENDING | DESCENDING`

Required: No

StatusEquals

A filter that retrieves only jobs with a specific status.

Type: String

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "NextToken": "string",
   "VectorEnrichmentJobSummaries": [
      {
         "Arn": "string",
         "CreationTime": "string",
         "DurationInSeconds": number,
         "Name": "string",
```

```
            "Status": "string",
            "Tags": {
               "string" : "string"
            },
            "Type": "string"
         }
      ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

VectorEnrichmentJobSummaries

Contains summary information about the Vector Enrichment jobs.

Type: Array of ListVectorEnrichmentJobOutputConfig objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

SearchRasterDataCollection

Service: Amazon SageMaker geospatial capabilities

Allows you run image query on a specific raster data collection to get a list of the satellite imagery matching the selected filters.

Request Syntax

```
POST /search-raster-data-collection HTTP/1.1
Content-type: application/json

{
   "Arn": "string",
   "NextToken": "string",
   "RasterDataCollectionQuery": {
      "AreaOfInterest": { ... },
      "BandFilter": [ "string" ],
      "PropertyFilters": {
         "LogicalOperator": "string",
         "Properties": [
            {
               "Property": { ... }
            }
         ]
      },
      "TimeRangeFilter": {
         "EndTime": number,
         "StartTime": number
      }
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

Arn

The Amazon Resource Name (ARN) of the raster data collection.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:raster-data-collection/(public|premium|user)/[a-z0-9]{12,}`

Required: Yes

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Required: No

RasterDataCollectionQuery

RasterDataCollectionQuery consisting of AreaOfInterest(AOI), PropertyFilters and TimeRangeFilterInput used in SearchRasterDataCollection.

Type: RasterDataCollectionQueryWithBandFilterInput object

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "ApproximateResultCount": number,
   "Items": [
      {
         "Assets": {
            "string" : {
               "Href": "string"
            }
         },
         "DateTime": number,
         "Geometry": {
```

```
            "Coordinates": [
               [
                  [ number ]
               ]
            ],
            "Type": "string"
         },
         "Id": "string",
         "Properties": {
            "EoCloudCover": number,
            "LandsatCloudCoverLand": number,
            "Platform": "string",
            "ViewOffNadir": number,
            "ViewSunAzimuth": number,
            "ViewSunElevation": number
         }
      }
   ],
   "NextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ApproximateResultCount

Approximate number of results in the response.

Type: Integer

Items

List of items matching the Raster DataCollectionQuery.

Type: Array of ItemSource objects

NextToken

If the previous response was truncated, you receive this token. Use it in your next request to receive the next set of results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8192.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StartEarthObservationJob

Service: Amazon SageMaker geospatial capabilities

Use this operation to create an Earth observation job.

Request Syntax

```
POST /earth-observation-jobs HTTP/1.1
Content-type: application/json

{
   "ClientToken": "string",
   "ExecutionRoleArn": "string",
   "InputConfig": {
      "PreviousEarthObservationJobArn": "string",
      "RasterDataCollectionQuery": {
         "AreaOfInterest": { ... },
         "PropertyFilters": {
            "LogicalOperator": "string",
            "Properties": [
               {
                  "Property": { ... }
               }
            ]
         },
         "RasterDataCollectionArn": "string",
         "TimeRangeFilter": {
            "EndTime": number,
            "StartTime": number
         }
      }
   },
   "JobConfig": { ... },
   "KmsKeyId": "string",
   "Name": "string",
   "Tags": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

ClientToken

A unique token that guarantees that the call to this API is idempotent.

Type: String

Length Constraints: Minimum length of 36. Maximum length of 64.

Required: No

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you specified for the job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:(aws[a-z-]*):iam::([0-9]{12}):role/[a-zA-Z0-9+=,.@_/-]+`

Required: Yes

InputConfig

Input configuration information for the Earth Observation job.

Type: InputConfigInput object

Required: Yes

JobConfig

An object containing information about the job configuration.

Type: JobConfigInput object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

KmsKeyId

The Key Management Service key ID for server-side encryption.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

Name

The name of the Earth Observation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 200.

Required: Yes

Tags

Each tag consists of a key and a value.

Type: String to string map

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "Arn": "string",
   "CreationTime": "string",
   "DurationInSeconds": number,
   "ExecutionRoleArn": "string",
   "InputConfig": {
      "PreviousEarthObservationJobArn": "string",
      "RasterDataCollectionQuery": {
         "AreaOfInterest": { ... },
         "PropertyFilters": {
            "LogicalOperator": "string",
            "Properties": [
               {
                  "Property": { ... }
               }
```

```
            ]
         },
         "RasterDataCollectionArn": "string",
         "RasterDataCollectionName": "string",
         "TimeRangeFilter": {
            "EndTime": "string",
            "StartTime": "string"
         }
      }
   },
   "JobConfig": { ... },
   "KmsKeyId": "string",
   "Name": "string",
   "Status": "string",
   "Tags": {
      "string" : "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The Amazon Resource Name (ARN) of the Earth Observation job.

Type: String

CreationTime

The creation time.

Type: Timestamp

DurationInSeconds

The duration of the session, in seconds.

Type: Integer

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you specified for the job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:(aws[a-z-]*):iam::([0-9]{12}):role/[a-zA-Z0-9+=,.@_/-]+`

InputConfig

Input configuration information for the Earth Observation job.

Type: InputConfigOutput object

JobConfig

An object containing information about the job configuration.

Type: JobConfigInput object

Note: This object is a Union. Only one member of this object can be specified or returned.

KmsKeyId

The Key Management Service key ID for server-side encryption.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Name

The name of the Earth Observation job.

Type: String

Status

The status of the Earth Observation job.

Type: String

Valid Values: `INITIALIZING | IN_PROGRESS | STOPPING | COMPLETED | STOPPED | FAILED | DELETING | DELETED`

Tags

Each tag consists of a key and a value.

Type: String to string map

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

ConflictException

Updating or deleting a resource can cause an inconsistent state.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 409

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ServiceQuotaExceededException

You have exceeded the service quota.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 402

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StartVectorEnrichmentJob

Service: Amazon SageMaker geospatial capabilities

Creates a Vector Enrichment job for the supplied job type. Currently, there are two supported job types: reverse geocoding and map matching.

Request Syntax

```
POST /vector-enrichment-jobs HTTP/1.1
Content-type: application/json

{
   "ClientToken": "string",
   "ExecutionRoleArn": "string",
   "InputConfig": {
      "DataSourceConfig": { ... },
      "DocumentType": "string"
   },
   "JobConfig": { ... },
   "KmsKeyId": "string",
   "Name": "string",
   "Tags": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

ClientToken

A unique token that guarantees that the call to this API is idempotent.

Type: String

Length Constraints: Minimum length of 36. Maximum length of 64.

Required: No

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you specified for the job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:(aws[a-z-]*):iam::([0-9]{12}):role/[a-zA-Z0-9+=,.@_/-]+`

Required: Yes

InputConfig

Input configuration information for the Vector Enrichment job.

Type: VectorEnrichmentJobInputConfig object

Required: Yes

JobConfig

An object containing information about the job configuration.

Type: VectorEnrichmentJobConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

KmsKeyId

The Key Management Service key ID for server-side encryption.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

Name

The name of the Vector Enrichment job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 200.

Required: Yes

Tags

Each tag consists of a key and a value.

Type: String to string map

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "Arn": "string",
   "CreationTime": "string",
   "DurationInSeconds": number,
   "ExecutionRoleArn": "string",
   "InputConfig": {
      "DataSourceConfig": { ... },
      "DocumentType": "string"
   },
   "JobConfig": { ... },
   "KmsKeyId": "string",
   "Name": "string",
   "Status": "string",
   "Tags": {
      "string" : "string"
   },
   "Type": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Arn

The Amazon Resource Name (ARN) of the Vector Enrichment job.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:vector-enrichment-job/[a-z0-9]{12,}`

CreationTime

The creation time.

Type: Timestamp

DurationInSeconds

The duration of the Vector Enrichment job, in seconds.

Type: Integer

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you specified for the job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:(aws[a-z-]*):iam::([0-9]{12}):role/[a-zA-Z0-9+=,.@_/-]+`

InputConfig

Input configuration information for starting the Vector Enrichment job.

Type: VectorEnrichmentJobInputConfig object

JobConfig

An object containing information about the job configuration.

Type: VectorEnrichmentJobConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

KmsKeyId

The Key Management Service key ID for server-side encryption.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Name

The name of the Vector Enrichment job.

Type: String

Status

The status of the Vector Enrichment job being started.

Type: String

Valid Values: `INITIALIZING | IN_PROGRESS | STOPPING | STOPPED | COMPLETED | FAILED | DELETING | DELETED`

Tags

Each tag consists of a key and a value.

Type: String to string map

Type

The type of the Vector Enrichment job.

Type: String

Valid Values: `REVERSE_GEOCODING | MAP_MATCHING`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

ConflictException

Updating or deleting a resource can cause an inconsistent state.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 409

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ServiceQuotaExceededException

You have exceeded the service quota.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 402

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopEarthObservationJob

Service: Amazon SageMaker geospatial capabilities

Use this operation to stop an existing earth observation job.

Request Syntax

```
POST /earth-observation-jobs/stop HTTP/1.1
Content-type: application/json

{
   "Arn": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

Arn

The Amazon Resource Name (ARN) of the Earth Observation job being stopped.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:earth-observation-job/[a-z0-9]{12,}`

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

ConflictException

Updating or deleting a resource can cause an inconsistent state.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 409

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopVectorEnrichmentJob

Service: Amazon SageMaker geospatial capabilities

Stops the Vector Enrichment job for a given job ARN.

Request Syntax

```
POST /vector-enrichment-jobs/stop HTTP/1.1
Content-type: application/json

{
   "Arn": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

Arn

The Amazon Resource Name (ARN) of the Vector Enrichment job.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:vector-enrichment-job/[a-z0-9]{12,}`

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

ConflictException

Updating or deleting a resource can cause an inconsistent state.

ResourceId

Identifier of the resource affected.

HTTP Status Code: 409

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

TagResource

Service: Amazon SageMaker geospatial capabilities

The resource you want to tag.

Request Syntax

```
PUT /tags/ResourceArn HTTP/1.1
Content-type: application/json

{
   "Tags": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

ResourceArn

The Amazon Resource Name (ARN) of the resource you want to tag.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Required: Yes

Request Body

The request accepts the following data in JSON format.

Tags

Each tag consists of a key and a value.

Type: String to string map

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UntagResource

Service: Amazon SageMaker geospatial capabilities

The resource you want to untag.

Request Syntax

```
DELETE /tags/ResourceArn?tagKeys=TagKeys HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

ResourceArn

The Amazon Resource Name (ARN) of the resource you want to untag.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Required: Yes

TagKeys

Keys of the tags you want to remove.

Array Members: Minimum number of 1 item. Maximum number of 50 items.

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 403

InternalServerException

The request processing has failed because of an unknown error, exception, or failure.

ResourceId

HTTP Status Code: 500

ResourceNotFoundException

The request references a resource which does not exist.

ResourceId

Identifier of the resource that was not found.

HTTP Status Code: 404

ThrottlingException

The request was denied due to request throttling.

ResourceId

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by an AWS service.

ResourceId

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Amazon SageMaker Metrics Service

The following actions are supported by Amazon SageMaker Metrics Service:

- BatchGetMetrics
- BatchPutMetrics

BatchGetMetrics

Service: Amazon SageMaker Metrics Service

Used to retrieve training metrics from SageMaker.

Request Syntax

```
POST /BatchGetMetrics HTTP/1.1
Content-type: application/json

{
   "MetricQueries": [
      {
         "End": number,
         "MetricName": "string",
         "MetricStat": "string",
         "Period": "string",
         "ResourceArn": "string",
         "Start": number,
         "XAxisType": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

MetricQueries

Queries made to retrieve training metrics from SageMaker.

Type: Array of MetricQuery objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "MetricQueryResults": [
      {
         "Message": "string",
         "MetricValues": [ number ],
         "Status": "string",
         "XAxisValues": [ number ]
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

MetricQueryResults

The results of a query to retrieve training metrics from SageMaker.

Type: Array of MetricQueryResult objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

BatchPutMetrics

Service: Amazon SageMaker Metrics Service

Used to ingest training metrics into SageMaker. These metrics can be visualized in SageMaker Studio.

Request Syntax

```
PUT /BatchPutMetrics HTTP/1.1
Content-type: application/json

{
   "MetricData": [
      {
         "MetricName": "string",
         "Step": number,
         "Timestamp": number,
         "Value": number
      }
   ],
   "TrialComponentName": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

MetricData

A list of raw metric values to put.

Type: Array of RawMetricData objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: Yes

TrialComponentName

The name of the Trial Component to associate with the metrics. The Trial Component name must be entirely lowercase.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `^[a-z0-9](-*[a-z0-9]){0,119}`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "Errors": [
      {
         "Code": "string",
         "MetricIndex": number
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Errors

Lists any errors that occur when inserting metric data.

Type: Array of BatchPutMetricsError objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Amazon SageMaker Job Runtime Service

The following actions are supported by Amazon SageMaker Job Runtime Service:

- CompleteRollout
- Sample
- SampleWithResponseStream
- UpdateReward

CompleteRollout

Service: Amazon SageMaker Job Runtime Service

Marks a rollout as complete, indicating that no further turns will be appended to the trajectory. After calling this operation, the trajectory is sealed and eligible for reward submission via the UpdateReward operation.

Request Syntax

```
POST /complete-rollout HTTP/1.1
X-Amzn-SageMaker-Job-Arn: JobArn
Content-type: application/json

{
   "ClientToken": "string",
   "Status": "string",
   "TrajectoryId": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

JobArn

The job ARN.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:job/[a-zA-Z0-9_\-]+/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

ClientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request.

Type: String

Required: No

Status

The target status for the trajectory. Defaults to READY if not specified. Set to FAILED if the rollout encountered an error and the trajectory should not be used for processing.

Type: String

Valid Values: `ready | failed`

Required: No

TrajectoryId

The trajectory ID to mark as complete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have permission to perform this operation.

HTTP Status Code: 403

ConflictException

The request conflicts with the current state of the resource.

HTTP Status Code: 409

InternalServiceError

An internal service error occurred. Retry the request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource was not found.

HTTP Status Code: 404

ServiceQuotaExceededException

You have exceeded a service quota.

HTTP Status Code: 402

ThrottlingException

The request was throttled. Retry the request after a brief wait.

HTTP Status Code: 429

ValidationException

The request is not valid. Check the request syntax and parameters.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Sample

Service: Amazon SageMaker Job Runtime Service

Sends an inference request to the model during a job execution. The request and response bodies are forwarded to and from the model without modification. Each turn (prompt and response) is captured for later use.

Request Syntax

```
POST /sample HTTP/1.1
X-Amzn-SageMaker-Job-Arn: JobArn
X-Amzn-SageMaker-Trajectory-Id: TrajectoryId

Body
```

URI Request Parameters

The request uses the following URI parameters.

JobArn

The job ARN that identifies which model session to route the inference request to.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:job/[a-zA-Z0-9_\-]+/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

TrajectoryId

The trajectory ID for grouping turns into a single rollout. Each turn (prompt and response) is captured for later use.

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: Yes

Request Body

The request accepts the following binary data.

Body

The raw inference request body in OpenAI-compatible JSON format.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-Type: ContentType

Body
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

ContentType

MIME type of the inference result.

The response returns the following as the HTTP body.

Body

The raw inference response body from the model.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have permission to perform this operation.

HTTP Status Code: 403

InternalServiceError

An internal service error occurred. Retry the request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource was not found.

HTTP Status Code: 404

ServiceQuotaExceededException

You have exceeded a service quota.

HTTP Status Code: 402

ThrottlingException

The request was throttled. Retry the request after a brief wait.

HTTP Status Code: 429

ValidationException

The request is not valid. Check the request syntax and parameters.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

SampleWithResponseStream

Service: Amazon SageMaker Job Runtime Service

Sends a streaming inference request to the model during a job execution. Returns the response as a stream of payload chunks. Each turn is captured for later use.

Request Syntax

```
POST /sample-with-response-stream HTTP/1.1
X-Amzn-SageMaker-Job-Arn: JobArn
X-Amzn-SageMaker-Trajectory-Id: TrajectoryId

Body
```

URI Request Parameters

The request uses the following URI parameters.

JobArn

The job ARN that identifies which model session to route the inference request to.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:job/[a-zA-Z0-9_\-]+/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

TrajectoryId

The trajectory ID for grouping turns into a single rollout. Each turn is captured for later use.

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: Yes

Request Body

The request accepts the following binary data.

Body

The raw inference request body in OpenAI-compatible JSON format.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-Type: ContentType

Body
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

ContentType

MIME type of the streaming inference result.

The response returns the following as the HTTP body.

Body

The streaming response body, delivered as a series of PayloadPart events.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have permission to perform this operation.

HTTP Status Code: 403

InternalServiceError

An internal service error occurred. Retry the request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource was not found.

HTTP Status Code: 404

ServiceQuotaExceededException

You have exceeded a service quota.

HTTP Status Code: 402

ThrottlingException

The request was throttled. Retry the request after a brief wait.

HTTP Status Code: 429

ValidationException

The request is not valid. Check the request syntax and parameters.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateReward

Service: Amazon SageMaker Job Runtime Service

Updates the reward values for a trajectory and transitions it to reward-received status, signaling that it is eligible for processing. Call this operation after CompleteRollout to provide the computed reward scores.

Request Syntax

```
POST /update-reward HTTP/1.1
X-Amzn-SageMaker-Job-Arn: JobArn
Content-type: application/json

{
   "ClientToken": "string",
   "Rewards": [ number ],
   "TrajectoryId": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

JobArn

The job ARN.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:job/[a-zA-Z0-9_\-]+/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

ClientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request.

Type: String

Required: No

Rewards

The list of reward values to assign to this trajectory. Provide one reward value per turn in the trajectory.

Type: Array of doubles

Required: Yes

TrajectoryId

The trajectory ID to update with reward values.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You do not have permission to perform this operation.

HTTP Status Code: 403

ConflictException

The request conflicts with the current state of the resource.

HTTP Status Code: 409

InternalServiceError

An internal service error occurred. Retry the request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource was not found.

HTTP Status Code: 404

ServiceQuotaExceededException

You have exceeded a service quota.

HTTP Status Code: 402

ThrottlingException

The request was throttled. Retry the request after a brief wait.

HTTP Status Code: 429

ValidationException

The request is not valid. Check the request syntax and parameters.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Data Types

The following data types are supported by Amazon SageMaker Service:

- AcceleratorPartitionConfig
- ActionSource
- ActionSummary
- AddClusterNodeSpecification
- AdditionalEnis
- AdditionalInferenceSpecificationDefinition
- AdditionalModelDataSource
- AdditionalS3DataSource
- AgentVersion
- AIBenchmarkEndpoint
- AIBenchmarkInferenceComponent
- AIBenchmarkJobSummary
- AIBenchmarkNetworkConfig
- AIBenchmarkOutputConfig
- AIBenchmarkOutputResult
- AIBenchmarkTarget
- AICapacityReservationConfig
- AICloudWatchLogs
- AIDatasetConfig
- AIMlflowConfig
- AIModelSource
- AIModelSourceS3
- AIRecommendation
- AIRecommendationComputeSpec
- AIRecommendationConstraint
- AIRecommendationDeploymentConfiguration
- AIRecommendationDeploymentS3Channel

- AIRecommendationInferenceSpecification
- AIRecommendationInstanceDetail
- AIRecommendationJobSummary
- AIRecommendationModelDetails
- AIRecommendationOptimizationDetail
- AIRecommendationOutputConfig
- AIRecommendationOutputResult
- AIRecommendationPerformanceMetric
- AIRecommendationPerformanceTarget
- AIWorkloadConfigs
- AIWorkloadConfigSummary
- AIWorkloadDataSource
- AIWorkloadInputDataConfig
- AIWorkloadS3DataSource
- Alarm
- AlarmDetails
- AlgorithmSpecification
- AlgorithmStatusDetails
- AlgorithmStatusItem
- AlgorithmSummary
- AlgorithmValidationProfile
- AlgorithmValidationSpecification
- AmazonQSettings
- AnnotationConsolidationConfig
- AppDetails
- AppImageConfigDetails
- AppLifecycleManagement
- AppSpecification
- ArtifactSource
- ArtifactSourceType

- ArtifactSummary
- AssociationInfo
- AssociationSummary
- AsyncInferenceClientConfig
- AsyncInferenceConfig
- AsyncInferenceNotificationConfig
- AsyncInferenceOutputConfig
- AthenaDatasetDefinition
- AuthorizedUrl
- AutoMLAlgorithmConfig
- AutoMLCandidate
- AutoMLCandidateGenerationConfig
- AutoMLCandidateStep
- AutoMLChannel
- AutoMLComputeConfig
- AutoMLContainerDefinition
- AutoMLDataSource
- AutoMLDataSplitConfig
- AutoMLJobArtifacts
- AutoMLJobChannel
- AutoMLJobCompletionCriteria
- AutoMLJobConfig
- AutoMLJobObjective
- AutoMLJobStepMetadata
- AutoMLJobSummary
- AutoMLOutputDataConfig
- AutoMLPartialFailureReason
- AutoMLProblemTypeConfig
- AutoMLProblemTypeResolvedAttributes
- AutoMLResolvedAttributes

- AutoMLS3DataSource
- AutoMLSecurityConfig
- AutoParameter
- AutoRollbackConfig
- Autotune
- AvailableUpgrade
- BaseModel
- BatchAddClusterNodesError
- BatchDataCaptureConfig
- BatchDeleteClusterNodeLogicalIdsError
- BatchDeleteClusterNodesError
- BatchDescribeModelPackageError
- BatchDescribeModelPackageSummary
- BatchRebootClusterNodeLogicalIdsError
- BatchRebootClusterNodesError
- BatchReplaceClusterNodeLogicalIdsError
- BatchReplaceClusterNodesError
- BatchTransformInput
- BedrockCustomModelDeploymentMetadata
- BedrockCustomModelMetadata
- BedrockModelImportMetadata
- BedrockProvisionedModelThroughputMetadata
- BestObjectiveNotImproving
- Bias
- BlueGreenUpdatePolicy
- CacheHitResult
- CallbackStepMetadata
- CandidateArtifactLocations
- CandidateGenerationConfig
- CandidateProperties

- CanvasAppSettings
- CapacityReservation
- CapacitySize
- CapacitySizeConfig
- CaptureContentTypeHeader
- CaptureOption
- CategoricalParameter
- CategoricalParameterRange
- CategoricalParameterRangeSpecification
- CfnCreateTemplateProvider
- CfnStackCreateParameter
- CfnStackDetail
- CfnStackParameter
- CfnStackUpdateParameter
- CfnTemplateProviderDetail
- CfnUpdateTemplateProvider
- Channel
- ChannelSpecification
- CheckpointConfig
- ClarifyCheckStepMetadata
- ClarifyExplainerConfig
- ClarifyInferenceConfig
- ClarifyShapBaselineConfig
- ClarifyShapConfig
- ClarifyTextConfig
- ClusterAutoScalingConfig
- ClusterAutoScalingConfigOutput
- ClusterCapacityRequirements
- ClusterEbsVolumeConfig
- ClusterEventDetail

- ClusterEventSummary
- ClusterFsxLustreConfig
- ClusterFsxOpenZfsConfig
- ClusterInstanceGroupDetails
- ClusterInstanceGroupSpecification
- ClusterInstancePlacement
- ClusterInstanceRequirementDetails
- ClusterInstanceRequirements
- ClusterInstanceStatusDetails
- ClusterInstanceStorageConfig
- ClusterInstanceTypeDetail
- ClusterKubernetesConfig
- ClusterKubernetesConfigDetails
- ClusterKubernetesConfigNodeDetails
- ClusterKubernetesTaint
- ClusterLifeCycleConfig
- ClusterMetadata
- ClusterNetworkInterface
- ClusterNetworkInterfaceDetails
- ClusterNodeDetails
- ClusterNodeSummary
- ClusterOnDemandOptions
- ClusterOrchestrator
- ClusterOrchestratorEksConfig
- ClusterOrchestratorSlurmConfig
- ClusterRestrictedInstanceGroupDetails
- ClusterRestrictedInstanceGroupsConfig
- ClusterRestrictedInstanceGroupsConfigOutput
- ClusterRestrictedInstanceGroupSpecification
- ClusterSchedulerConfigSummary

- ClusterSharedEnvironmentConfig
- ClusterSharedEnvironmentConfigDetails
- ClusterSlurmConfig
- ClusterSlurmConfigDetails
- ClusterSpotOptions
- ClusterSummary
- ClusterTieredStorageConfig
- CodeEditorAppImageConfig
- CodeEditorAppSettings
- CodeRepository
- CodeRepositorySummary
- CognitoConfig
- CognitoMemberDefinition
- CollectionConfig
- CollectionConfiguration
- CompilationJobSummary
- ComputeQuotaConfig
- ComputeQuotaResourceConfig
- ComputeQuotaSummary
- ComputeQuotaTarget
- ConditionStepMetadata
- ContainerConfig
- ContainerDefinition
- ContainerMetricsConfig
- ContextSource
- ContextSummary
- ContinuousParameterRange
- ContinuousParameterRangeSpecification
- ConvergenceDetected
- CreateTemplateProvider

- CustomFileSystem
- CustomFileSystemConfig
- CustomImage
- CustomizedMetricSpecification
- CustomPosixUserConfig
- DataCaptureConfig
- DataCaptureConfigSummary
- DataCatalogConfig
- DataProcessing
- DataQualityAppSpecification
- DataQualityBaselineConfig
- DataQualityJobInput
- DatasetDefinition
- DatasetSource
- DataSource
- DebugHookConfig
- DebugRuleConfiguration
- DebugRuleEvaluationStatus
- DefaultEbsStorageSettings
- DefaultSpaceSettings
- DefaultSpaceStorageSettings
- DeployedImage
- DeploymentConfig
- DeploymentConfiguration
- DeploymentRecommendation
- DeploymentStage
- DeploymentStageStatusSummary
- DerivedInformation
- DesiredWeightAndCapacity
- Device

- DeviceDeploymentSummary
- DeviceFleetSummary
- DeviceSelectionConfig
- DeviceStats
- DeviceSummary
- DirectDeploySettings
- DockerSettings
- DomainDetails
- DomainSettings
- DomainSettingsForUpdate
- DriftCheckBaselines
- DriftCheckBias
- DriftCheckExplainability
- DriftCheckModelDataQuality
- DriftCheckModelQuality
- DynamicScalingConfiguration
- EbsStorageSettings
- Ec2CapacityReservation
- Edge
- EdgeDeploymentConfig
- EdgeDeploymentModelConfig
- EdgeDeploymentPlanSummary
- EdgeDeploymentStatus
- EdgeModel
- EdgeModelStat
- EdgeModelSummary
- EdgeOutputConfig
- EdgePackagingJobSummary
- EdgePresetDeploymentOutput
- EFSFileSystem

- EFSFileSystemConfig
- EmrServerlessComputeConfig
- EmrServerlessSettings
- EmrSettings
- EMRStepMetadata
- Endpoint
- EndpointConfigStepMetadata
- EndpointConfigSummary
- EndpointInfo
- EndpointInput
- EndpointInputConfiguration
- EndpointMetadata
- EndpointOutputConfiguration
- EndpointPerformance
- EndpointStepMetadata
- EndpointSummary
- EnvironmentConfig
- EnvironmentConfigDetails
- EnvironmentParameter
- EnvironmentParameterRanges
- ErrorInfo
- EventDetails
- EventMetadata
- Experiment
- ExperimentConfig
- ExperimentSource
- ExperimentSummary
- Explainability
- ExplainerConfig
- FailStepMetadata

- FeatureDefinition
- FeatureGroup
- FeatureGroupSummary
- FeatureMetadata
- FeatureParameter
- FileSource
- FileSystemConfig
- FileSystemDataSource
- Filter
- FinalAutoMLJobObjectiveMetric
- FinalHyperParameterTuningJobObjectiveMetric
- FlowDefinitionOutputConfig
- FlowDefinitionSummary
- FSxLustreConfig
- FSxLustreFileSystem
- FSxLustreFileSystemConfig
- GenerativeAiSettings
- GitConfig
- GitConfigForUpdate
- HiddenSageMakerImage
- HolidayConfigAttributes
- HubAccessConfig
- HubContentDependency
- HubContentInfo
- HubInfo
- HubS3StorageConfig
- HumanLoopActivationConditionsConfig
- HumanLoopActivationConfig
- HumanLoopConfig
- HumanLoopRequestSource

- HumanTaskConfig
- HumanTaskUiSummary
- HyperbandStrategyConfig
- HyperParameterAlgorithmSpecification
- HyperParameterSpecification
- HyperParameterTrainingJobDefinition
- HyperParameterTrainingJobSummary
- HyperParameterTuningInstanceConfig
- HyperParameterTuningJobCompletionDetails
- HyperParameterTuningJobConfig
- HyperParameterTuningJobConsumedResources
- HyperParameterTuningJobObjective
- HyperParameterTuningJobSearchEntity
- HyperParameterTuningJobStrategyConfig
- HyperParameterTuningJobSummary
- HyperParameterTuningJobWarmStartConfig
- HyperParameterTuningResourceConfig
- IamIdentity
- IamPolicyConstraints
- IdentityProviderOAuthSetting
- IdleSettings
- Image
- ImageClassificationJobConfig
- ImageConfig
- ImageVersion
- InferenceComponentAvailabilityZoneBalance
- InferenceComponentCapacitySize
- InferenceComponentComputeResourceRequirements
- InferenceComponentContainerSpecification
- InferenceComponentContainerSpecificationSummary

- InferenceComponentDataCacheConfig
- InferenceComponentDataCacheConfigSummary
- InferenceComponentDeploymentConfig
- InferenceComponentMetadata
- InferenceComponentPlacementStatus
- InferenceComponentRollingUpdatePolicy
- InferenceComponentRuntimeConfig
- InferenceComponentRuntimeConfigSummary
- InferenceComponentSchedulingConfig
- InferenceComponentSpecification
- InferenceComponentSpecificationSummary
- InferenceComponentStartupParameters
- InferenceComponentSummary
- InferenceExecutionConfig
- InferenceExperimentDataStorageConfig
- InferenceExperimentSchedule
- InferenceExperimentSummary
- InferenceHubAccessConfig
- InferenceMetrics
- InferenceRecommendation
- InferenceRecommendationsJob
- InferenceRecommendationsJobStep
- InferenceSpecification
- InfraCheckConfig
- InputConfig
- InstanceGroup
- InstanceGroupHealthCheckConfiguration
- InstanceGroupMetadata
- InstanceGroupScalingMetadata
- InstanceMetadata

- InstanceMetadataServiceConfiguration
- InstancePlacementConfig
- InstancePool
- InstancePoolSummary
- InstanceRequirementsEniConfiguration
- IntegerParameterRange
- IntegerParameterRangeSpecification
- Job
- JobConfigSchemaVersionSummary
- JobSecondaryStatusTransition
- JobStepMetadata
- JobSummary
- JupyterLabAppImageConfig
- JupyterLabAppSettings
- JupyterServerAppSettings
- KendraSettings
- KernelGatewayAppSettings
- KernelGatewayImageConfig
- KernelSpec
- LabelCounters
- LabelCountersForWorkteam
- LabelingJobAlgorithmsConfig
- LabelingJobDataAttributes
- LabelingJobDataSource
- LabelingJobForWorkteamSummary
- LabelingJobInputConfig
- LabelingJobOutput
- LabelingJobOutputConfig
- LabelingJobResourceConfig
- LabelingJobS3DataSource

- LabelingJobSnsDataSource
- LabelingJobStoppingConditions
- LabelingJobSummary
- LambdaStepMetadata
- LastUpdateStatus
- LineageGroupSummary
- LineageMetadata
- ManagedConfiguration
- MemberDefinition
- MetadataProperties
- MetricData
- MetricDatum
- MetricDefinition
- MetricsConfig
- MetricsEndpoint
- MetricSpecification
- MetricsSource
- MlflowAppSummary
- MlflowConfig
- MLflowConfiguration
- MlflowDetails
- Model
- ModelAccessConfig
- ModelArtifacts
- ModelBiasAppSpecification
- ModelBiasBaselineConfig
- ModelBiasJobInput
- ModelCard
- ModelCardExportArtifacts
- ModelCardExportJobSummary

- ModelCardExportOutputConfig
- ModelCardSecurityConfig
- ModelCardSummary
- ModelCardVersionSummary
- ModelClientConfig
- ModelCompilationConfig
- ModelConfiguration
- ModelDashboardEndpoint
- ModelDashboardIndicatorAction
- ModelDashboardModel
- ModelDashboardModelCard
- ModelDashboardMonitoringSchedule
- ModelDataQuality
- ModelDataSource
- ModelDeployConfig
- ModelDeployResult
- ModelDigests
- ModelExplainabilityAppSpecification
- ModelExplainabilityBaselineConfig
- ModelExplainabilityJobInput
- ModelInfrastructureConfig
- ModelInput
- ModelLatencyThreshold
- ModelLifeCycle
- ModelMetadataFilter
- ModelMetadataSearchExpression
- ModelMetadataSummary
- ModelMetrics
- ModelPackage
- ModelPackageConfig

- ModelPackageContainerDefinition
- ModelPackageGroup
- ModelPackageGroupSummary
- ModelPackageModelCard
- ModelPackageSecurityConfig
- ModelPackageStatusDetails
- ModelPackageStatusItem
- ModelPackageSummary
- ModelPackageValidationProfile
- ModelPackageValidationSpecification
- ModelQuality
- ModelQualityAppSpecification
- ModelQualityBaselineConfig
- ModelQualityJobInput
- ModelQuantizationConfig
- ModelRegisterSettings
- ModelShardingConfig
- ModelSpeculativeDecodingConfig
- ModelSpeculativeDecodingTrainingDataSource
- ModelStepMetadata
- ModelSummary
- ModelVariantConfig
- ModelVariantConfigSummary
- MonitoringAlertActions
- MonitoringAlertHistorySummary
- MonitoringAlertSummary
- MonitoringAppSpecification
- MonitoringBaselineConfig
- MonitoringClusterConfig
- MonitoringConstraintsResource

- MonitoringCsvDatasetFormat
- MonitoringDatasetFormat
- MonitoringExecutionSummary
- MonitoringGroundTruthS3Input
- MonitoringInput
- MonitoringJobDefinition
- MonitoringJobDefinitionSummary
- MonitoringJsonDatasetFormat
- MonitoringNetworkConfig
- MonitoringOutput
- MonitoringOutputConfig
- MonitoringParquetDatasetFormat
- MonitoringResources
- MonitoringS3Output
- MonitoringSchedule
- MonitoringScheduleConfig
- MonitoringScheduleSummary
- MonitoringStatisticsResource
- MonitoringStoppingCondition
- MultiModelConfig
- NeoVpcConfig
- NestedFilters
- NetworkConfig
- NodeAdditionResult
- NotebookInstanceLifecycleConfigSummary
- NotebookInstanceLifecycleHook
- NotebookInstanceSummary
- NotificationConfiguration
- ObjectiveStatusCounters
- OfflineStoreConfig

- OfflineStoreStatus
- OidcConfig
- OidcConfigForResponse
- OidcMemberDefinition
- OnlineStoreConfig
- OnlineStoreConfigUpdate
- OnlineStoreSecurityConfig
- OptimizationConfig
- OptimizationJobModelSource
- OptimizationJobModelSourceS3
- OptimizationJobOutputConfig
- OptimizationJobSummary
- OptimizationModelAccessConfig
- OptimizationOutput
- OptimizationSageMakerModel
- OptimizationVpcConfig
- OutputConfig
- OutputDataConfig
- OutputParameter
- OwnershipSettings
- OwnershipSettingsSummary
- ParallelismConfiguration
- Parameter
- ParameterRange
- ParameterRanges
- Parent
- ParentHyperParameterTuningJob
- PartnerAppConfig
- PartnerAppMaintenanceConfig
- PartnerAppSummary

- PendingDeploymentSummary
- PendingProductionVariantSummary
- Phase
- Pipeline
- PipelineDefinitionS3Location
- PipelineExecution
- PipelineExecutionStep
- PipelineExecutionStepMetadata
- PipelineExecutionSummary
- PipelineExperimentConfig
- PipelineSummary
- PipelineVersion
- PipelineVersionSummary
- PlacementSpecification
- PredefinedMetricSpecification
- PresignedUrlAccessConfig
- PriorityClass
- ProcessingClusterConfig
- ProcessingFeatureStoreOutput
- ProcessingInput
- ProcessingJob
- ProcessingJobStepMetadata
- ProcessingJobSummary
- ProcessingOutput
- ProcessingOutputConfig
- ProcessingResources
- ProcessingS3Input
- ProcessingS3Output
- ProcessingStoppingCondition
- ProductionVariant

- ProductionVariantCapacityReservationConfig
- ProductionVariantCapacityReservationSummary
- ProductionVariantCoreDumpConfig
- ProductionVariantManagedInstanceScaling
- ProductionVariantManagedInstanceScalingScaleInPolicy
- ProductionVariantRoutingConfig
- ProductionVariantServerlessConfig
- ProductionVariantServerlessUpdateConfig
- ProductionVariantStatus
- ProductionVariantSummary
- ProfilerConfig
- ProfilerConfigForUpdate
- ProfilerRuleConfiguration
- ProfilerRuleEvaluationStatus
- Project
- ProjectSummary
- PropertyNameQuery
- PropertyNameSuggestion
- ProvisioningParameter
- PublicWorkforceTaskPrice
- QualityCheckStepMetadata
- QueryFilters
- RealTimeInferenceConfig
- RealTimeInferenceRecommendation
- RecommendationJobCompiledOutputConfig
- RecommendationJobContainerConfig
- RecommendationJobInferenceBenchmark
- RecommendationJobInputConfig
- RecommendationJobOutputConfig
- RecommendationJobPayloadConfig

- RecommendationJobResourceLimit
- RecommendationJobStoppingConditions
- RecommendationJobVpcConfig
- RecommendationMetrics
- RedshiftDatasetDefinition
- RegisterModelStepMetadata
- RemoteDebugConfig
- RemoteDebugConfigForUpdate
- RenderableTask
- RenderingError
- RepositoryAuthConfig
- ReservedCapacityOffering
- ReservedCapacitySummary
- ResolvedAttributes
- ResourceCatalog
- ResourceConfig
- ResourceConfigForUpdate
- ResourceLimits
- ResourceSharingConfig
- ResourceSpec
- RetentionPolicy
- RetryStrategy
- RoleGroupAssignment
- RollingDeploymentPolicy
- RollingUpdatePolicy
- RSessionAppSettings
- RStudioServerProAppSettings
- RStudioServerProDomainSettings
- RStudioServerProDomainSettingsForUpdate
- S3DataSource

- S3FileSystem
- S3FileSystemConfig
- S3ModelDataSource
- S3Presign
- S3StorageConfig
- ScalingPolicy
- ScalingPolicyMetric
- ScalingPolicyObjective
- ScheduleConfig
- ScheduledUpdateConfig
- SchedulerConfig
- SearchExpression
- SearchRecord
- SecondaryStatusTransition
- SelectedStep
- SelectiveExecutionConfig
- SelectiveExecutionResult
- ServerlessJobConfig
- ServiceCatalogProvisionedProductDetails
- ServiceCatalogProvisioningDetails
- ServiceCatalogProvisioningUpdateDetails
- SessionChainingConfig
- ShadowModeConfig
- ShadowModelVariantConfig
- SharingSettings
- ShuffleConfig
- SourceAlgorithm
- SourceAlgorithmSpecification
- SourceIpConfig
- SpaceAppLifecycleManagement

- SpaceCodeEditorAppSettings
- SpaceDetails
- SpaceIdleSettings
- SpaceJupyterLabAppSettings
- SpaceSettings
- SpaceSettingsSummary
- SpaceSharingSettings
- SpaceSharingSettingsSummary
- SpaceStorageSettings
- Stairs
- StoppingCondition
- StudioLifecycleConfigDetails
- StudioWebPortalSettings
- SubscribedWorkteam
- SuggestionQuery
- TabularJobConfig
- TabularResolvedAttributes
- Tag
- TargetPlatform
- TargetTrackingScalingPolicyConfiguration
- TemplateProviderDetail
- TensorBoardAppSettings
- TensorBoardOutputConfig
- TextClassificationJobConfig
- TextGenerationJobConfig
- TextGenerationResolvedAttributes
- ThroughputConfig
- ThroughputConfigDescription
- ThroughputConfigUpdate
- TimeSeriesConfig

- TimeSeriesForecastingJobConfig
- TimeSeriesForecastingSettings
- TimeSeriesTransformations
- TotalHits
- TrackingServerSummary
- TrafficPattern
- TrafficRoutingConfig
- TrainingImageConfig
- TrainingJob
- TrainingJobDefinition
- TrainingJobStatusCounters
- TrainingJobStepMetadata
- TrainingJobSummary
- TrainingPlanExtension
- TrainingPlanExtensionOffering
- TrainingPlanFilter
- TrainingPlanOffering
- TrainingPlanSummary
- TrainingProgressInfo
- TrainingRepositoryAuthConfig
- TrainingSpecification
- TransformDataSource
- TransformInput
- TransformJob
- TransformJobDefinition
- TransformJobStepMetadata
- TransformJobSummary
- TransformOutput
- TransformResources
- TransformS3DataSource

- Trial
- TrialComponent
- TrialComponentArtifact
- TrialComponentMetricSummary
- TrialComponentParameterValue
- TrialComponentSimpleSummary
- TrialComponentSource
- TrialComponentSourceDetail
- TrialComponentStatus
- TrialComponentSummary
- TrialSource
- TrialSummary
- TrustedIdentityPropagationSettings
- TtlDuration
- TuningJobCompletionCriteria
- TuningJobStepMetaData
- UiConfig
- UiTemplate
- UiTemplateInfo
- UltraServer
- UltraServerInfo
- UltraServerSummary
- UnifiedStudioSettings
- UpdateClusterSoftwareInstanceGroupSpecification
- UpdateTemplateProvider
- USD
- UserContext
- UserProfileDetails
- UserSettings
- VariantProperty

- VectorConfig
- Vertex
- VisibilityConditions
- VpcConfig
- WarmPoolStatus
- WorkerAccessConfiguration
- Workforce
- WorkforceVpcConfigRequest
- WorkforceVpcConfigResponse
- WorkloadSpec
- WorkspaceSettings
- Workteam

The following data types are supported by Amazon SageMaker Runtime:

- PayloadPart
- ResponseStream

The following data types are supported by Amazon Sagemaker Edge Manager:

- Checksum
- Definition
- DeploymentModel
- DeploymentResult
- EdgeDeployment
- EdgeMetric
- Model

The following data types are supported by Amazon SageMaker Feature Store Runtime:

- BatchGetRecordError
- BatchGetRecordIdentifier

- BatchGetRecordResultDetail
- FeatureValue
- TtlDuration

The following data types are supported by Amazon SageMaker geospatial capabilities:

- AreaOfInterest
- AreaOfInterestGeometry
- AssetValue
- BandMathConfigInput
- CloudMaskingConfigInput
- CloudRemovalConfigInput
- CustomIndicesInput
- EarthObservationJobErrorDetails
- EoCloudCoverInput
- ExportErrorDetails
- ExportErrorDetailsOutput
- ExportS3DataInput
- ExportVectorEnrichmentJobOutputConfig
- Filter
- Geometry
- GeoMosaicConfigInput
- InputConfigInput
- InputConfigOutput
- ItemSource
- JobConfigInput
- LandCoverSegmentationConfigInput
- LandsatCloudCoverLandInput
- ListEarthObservationJobOutputConfig
- ListVectorEnrichmentJobOutputConfig
- MapMatchingConfig

- MultiPolygonGeometryInput
- Operation
- OutputBand
- OutputConfigInput
- OutputResolutionResamplingInput
- OutputResolutionStackInput
- PlatformInput
- PolygonGeometryInput
- Properties
- Property
- PropertyFilter
- PropertyFilters
- RasterDataCollectionMetadata
- RasterDataCollectionQueryInput
- RasterDataCollectionQueryOutput
- RasterDataCollectionQueryWithBandFilterInput
- ResamplingConfigInput
- ReverseGeocodingConfig
- StackConfigInput
- TemporalStatisticsConfigInput
- TimeRangeFilterInput
- TimeRangeFilterOutput
- UserDefined
- VectorEnrichmentJobConfig
- VectorEnrichmentJobDataSourceConfigInput
- VectorEnrichmentJobErrorDetails
- VectorEnrichmentJobExportErrorDetails
- VectorEnrichmentJobInputConfig
- VectorEnrichmentJobS3Data
- ViewOffNadirInput

- ViewSunAzimuthInput
- ViewSunElevationInput
- ZonalStatisticsConfigInput

The following data types are supported by Amazon SageMaker Metrics Service:

- BatchPutMetricsError
- MetricQuery
- MetricQueryResult
- RawMetricData

The following data types are supported by Amazon SageMaker Job Runtime Service:

Amazon SageMaker Service

The following data types are supported by Amazon SageMaker Service:

- AcceleratorPartitionConfig
- ActionSource
- ActionSummary
- AddClusterNodeSpecification
- AdditionalEnis
- AdditionalInferenceSpecificationDefinition
- AdditionalModelDataSource
- AdditionalS3DataSource
- AgentVersion
- AIBenchmarkEndpoint
- AIBenchmarkInferenceComponent
- AIBenchmarkJobSummary
- AIBenchmarkNetworkConfig
- AIBenchmarkOutputConfig
- AIBenchmarkOutputResult

- AIBenchmarkTarget
- AICapacityReservationConfig
- AICloudWatchLogs
- AIDatasetConfig
- AIMlflowConfig
- AIModelSource
- AIModelSourceS3
- AIRecommendation
- AIRecommendationComputeSpec
- AIRecommendationConstraint
- AIRecommendationDeploymentConfiguration
- AIRecommendationDeploymentS3Channel
- AIRecommendationInferenceSpecification
- AIRecommendationInstanceDetail
- AIRecommendationJobSummary
- AIRecommendationModelDetails
- AIRecommendationOptimizationDetail
- AIRecommendationOutputConfig
- AIRecommendationOutputResult
- AIRecommendationPerformanceMetric
- AIRecommendationPerformanceTarget
- AIWorkloadConfigs
- AIWorkloadConfigSummary
- AIWorkloadDataSource
- AIWorkloadInputDataConfig
- AIWorkloadS3DataSource
- Alarm
- AlarmDetails
- AlgorithmSpecification
- AlgorithmStatusDetails

- AlgorithmStatusItem
- AlgorithmSummary
- AlgorithmValidationProfile
- AlgorithmValidationSpecification
- AmazonQSettings
- AnnotationConsolidationConfig
- AppDetails
- AppImageConfigDetails
- AppLifecycleManagement
- AppSpecification
- ArtifactSource
- ArtifactSourceType
- ArtifactSummary
- AssociationInfo
- AssociationSummary
- AsyncInferenceClientConfig
- AsyncInferenceConfig
- AsyncInferenceNotificationConfig
- AsyncInferenceOutputConfig
- AthenaDatasetDefinition
- AuthorizedUrl
- AutoMLAlgorithmConfig
- AutoMLCandidate
- AutoMLCandidateGenerationConfig
- AutoMLCandidateStep
- AutoMLChannel
- AutoMLComputeConfig
- AutoMLContainerDefinition
- AutoMLDataSource
- AutoMLDataSplitConfig

- AutoMLJobArtifacts
- AutoMLJobChannel
- AutoMLJobCompletionCriteria
- AutoMLJobConfig
- AutoMLJobObjective
- AutoMLJobStepMetadata
- AutoMLJobSummary
- AutoMLOutputDataConfig
- AutoMLPartialFailureReason
- AutoMLProblemTypeConfig
- AutoMLProblemTypeResolvedAttributes
- AutoMLResolvedAttributes
- AutoMLS3DataSource
- AutoMLSecurityConfig
- AutoParameter
- AutoRollbackConfig
- Autotune
- AvailableUpgrade
- BaseModel
- BatchAddClusterNodesError
- BatchDataCaptureConfig
- BatchDeleteClusterNodeLogicalIdsError
- BatchDeleteClusterNodesError
- BatchDescribeModelPackageError
- BatchDescribeModelPackageSummary
- BatchRebootClusterNodeLogicalIdsError
- BatchRebootClusterNodesError
- BatchReplaceClusterNodeLogicalIdsError
- BatchReplaceClusterNodesError
- BatchTransformInput

- BedrockCustomModelDeploymentMetadata
- BedrockCustomModelMetadata
- BedrockModelImportMetadata
- BedrockProvisionedModelThroughputMetadata
- BestObjectiveNotImproving
- Bias
- BlueGreenUpdatePolicy
- CacheHitResult
- CallbackStepMetadata
- CandidateArtifactLocations
- CandidateGenerationConfig
- CandidateProperties
- CanvasAppSettings
- CapacityReservation
- CapacitySize
- CapacitySizeConfig
- CaptureContentTypeHeader
- CaptureOption
- CategoricalParameter
- CategoricalParameterRange
- CategoricalParameterRangeSpecification
- CfnCreateTemplateProvider
- CfnStackCreateParameter
- CfnStackDetail
- CfnStackParameter
- CfnStackUpdateParameter
- CfnTemplateProviderDetail
- CfnUpdateTemplateProvider
- Channel
- ChannelSpecification

- CheckpointConfig
- ClarifyCheckStepMetadata
- ClarifyExplainerConfig
- ClarifyInferenceConfig
- ClarifyShapBaselineConfig
- ClarifyShapConfig
- ClarifyTextConfig
- ClusterAutoScalingConfig
- ClusterAutoScalingConfigOutput
- ClusterCapacityRequirements
- ClusterEbsVolumeConfig
- ClusterEventDetail
- ClusterEventSummary
- ClusterFsxLustreConfig
- ClusterFsxOpenZfsConfig
- ClusterInstanceGroupDetails
- ClusterInstanceGroupSpecification
- ClusterInstancePlacement
- ClusterInstanceRequirementDetails
- ClusterInstanceRequirements
- ClusterInstanceStatusDetails
- ClusterInstanceStorageConfig
- ClusterInstanceTypeDetail
- ClusterKubernetesConfig
- ClusterKubernetesConfigDetails
- ClusterKubernetesConfigNodeDetails
- ClusterKubernetesTaint
- ClusterLifeCycleConfig
- ClusterMetadata
- ClusterNetworkInterface

- ClusterNetworkInterfaceDetails
- ClusterNodeDetails
- ClusterNodeSummary
- ClusterOnDemandOptions
- ClusterOrchestrator
- ClusterOrchestratorEksConfig
- ClusterOrchestratorSlurmConfig
- ClusterRestrictedInstanceGroupDetails
- ClusterRestrictedInstanceGroupsConfig
- ClusterRestrictedInstanceGroupsConfigOutput
- ClusterRestrictedInstanceGroupSpecification
- ClusterSchedulerConfigSummary
- ClusterSharedEnvironmentConfig
- ClusterSharedEnvironmentConfigDetails
- ClusterSlurmConfig
- ClusterSlurmConfigDetails
- ClusterSpotOptions
- ClusterSummary
- ClusterTieredStorageConfig
- CodeEditorAppImageConfig
- CodeEditorAppSettings
- CodeRepository
- CodeRepositorySummary
- CognitoConfig
- CognitoMemberDefinition
- CollectionConfig
- CollectionConfiguration
- CompilationJobSummary
- ComputeQuotaConfig
- ComputeQuotaResourceConfig

- ComputeQuotaSummary
- ComputeQuotaTarget
- ConditionStepMetadata
- ContainerConfig
- ContainerDefinition
- ContainerMetricsConfig
- ContextSource
- ContextSummary
- ContinuousParameterRange
- ContinuousParameterRangeSpecification
- ConvergenceDetected
- CreateTemplateProvider
- CustomFileSystem
- CustomFileSystemConfig
- CustomImage
- CustomizedMetricSpecification
- CustomPosixUserConfig
- DataCaptureConfig
- DataCaptureConfigSummary
- DataCatalogConfig
- DataProcessing
- DataQualityAppSpecification
- DataQualityBaselineConfig
- DataQualityJobInput
- DatasetDefinition
- DatasetSource
- DataSource
- DebugHookConfig
- DebugRuleConfiguration
- DebugRuleEvaluationStatus

- DefaultEbsStorageSettings
- DefaultSpaceSettings
- DefaultSpaceStorageSettings
- DeployedImage
- DeploymentConfig
- DeploymentConfiguration
- DeploymentRecommendation
- DeploymentStage
- DeploymentStageStatusSummary
- DerivedInformation
- DesiredWeightAndCapacity
- Device
- DeviceDeploymentSummary
- DeviceFleetSummary
- DeviceSelectionConfig
- DeviceStats
- DeviceSummary
- DirectDeploySettings
- DockerSettings
- DomainDetails
- DomainSettings
- DomainSettingsForUpdate
- DriftCheckBaselines
- DriftCheckBias
- DriftCheckExplainability
- DriftCheckModelDataQuality
- DriftCheckModelQuality
- DynamicScalingConfiguration
- EbsStorageSettings
- Ec2CapacityReservation

- Edge
- EdgeDeploymentConfig
- EdgeDeploymentModelConfig
- EdgeDeploymentPlanSummary
- EdgeDeploymentStatus
- EdgeModel
- EdgeModelStat
- EdgeModelSummary
- EdgeOutputConfig
- EdgePackagingJobSummary
- EdgePresetDeploymentOutput
- EFSFileSystem
- EFSFileSystemConfig
- EmrServerlessComputeConfig
- EmrServerlessSettings
- EmrSettings
- EMRStepMetadata
- Endpoint
- EndpointConfigStepMetadata
- EndpointConfigSummary
- EndpointInfo
- EndpointInput
- EndpointInputConfiguration
- EndpointMetadata
- EndpointOutputConfiguration
- EndpointPerformance
- EndpointStepMetadata
- EndpointSummary
- EnvironmentConfig
- EnvironmentConfigDetails

- EnvironmentParameter
- EnvironmentParameterRanges
- ErrorInfo
- EventDetails
- EventMetadata
- Experiment
- ExperimentConfig
- ExperimentSource
- ExperimentSummary
- Explainability
- ExplainerConfig
- FailStepMetadata
- FeatureDefinition
- FeatureGroup
- FeatureGroupSummary
- FeatureMetadata
- FeatureParameter
- FileSource
- FileSystemConfig
- FileSystemDataSource
- Filter
- FinalAutoMLJobObjectiveMetric
- FinalHyperParameterTuningJobObjectiveMetric
- FlowDefinitionOutputConfig
- FlowDefinitionSummary
- FSxLustreConfig
- FSxLustreFileSystem
- FSxLustreFileSystemConfig
- GenerativeAiSettings
- GitConfig

- GitConfigForUpdate
- HiddenSageMakerImage
- HolidayConfigAttributes
- HubAccessConfig
- HubContentDependency
- HubContentInfo
- HubInfo
- HubS3StorageConfig
- HumanLoopActivationConditionsConfig
- HumanLoopActivationConfig
- HumanLoopConfig
- HumanLoopRequestSource
- HumanTaskConfig
- HumanTaskUiSummary
- HyperbandStrategyConfig
- HyperParameterAlgorithmSpecification
- HyperParameterSpecification
- HyperParameterTrainingJobDefinition
- HyperParameterTrainingJobSummary
- HyperParameterTuningInstanceConfig
- HyperParameterTuningJobCompletionDetails
- HyperParameterTuningJobConfig
- HyperParameterTuningJobConsumedResources
- HyperParameterTuningJobObjective
- HyperParameterTuningJobSearchEntity
- HyperParameterTuningJobStrategyConfig
- HyperParameterTuningJobSummary
- HyperParameterTuningJobWarmStartConfig
- HyperParameterTuningResourceConfig
- IamIdentity

- IamPolicyConstraints
- IdentityProviderOAuthSetting
- IdleSettings
- Image
- ImageClassificationJobConfig
- ImageConfig
- ImageVersion
- InferenceComponentAvailabilityZoneBalance
- InferenceComponentCapacitySize
- InferenceComponentComputeResourceRequirements
- InferenceComponentContainerSpecification
- InferenceComponentContainerSpecificationSummary
- InferenceComponentDataCacheConfig
- InferenceComponentDataCacheConfigSummary
- InferenceComponentDeploymentConfig
- InferenceComponentMetadata
- InferenceComponentPlacementStatus
- InferenceComponentRollingUpdatePolicy
- InferenceComponentRuntimeConfig
- InferenceComponentRuntimeConfigSummary
- InferenceComponentSchedulingConfig
- InferenceComponentSpecification
- InferenceComponentSpecificationSummary
- InferenceComponentStartupParameters
- InferenceComponentSummary
- InferenceExecutionConfig
- InferenceExperimentDataStorageConfig
- InferenceExperimentSchedule
- InferenceExperimentSummary
- InferenceHubAccessConfig

- InferenceMetrics
- InferenceRecommendation
- InferenceRecommendationsJob
- InferenceRecommendationsJobStep
- InferenceSpecification
- InfraCheckConfig
- InputConfig
- InstanceGroup
- InstanceGroupHealthCheckConfiguration
- InstanceGroupMetadata
- InstanceGroupScalingMetadata
- InstanceMetadata
- InstanceMetadataServiceConfiguration
- InstancePlacementConfig
- InstancePool
- InstancePoolSummary
- InstanceRequirementsEniConfiguration
- IntegerParameterRange
- IntegerParameterRangeSpecification
- Job
- JobConfigSchemaVersionSummary
- JobSecondaryStatusTransition
- JobStepMetadata
- JobSummary
- JupyterLabAppImageConfig
- JupyterLabAppSettings
- JupyterServerAppSettings
- KendraSettings
- KernelGatewayAppSettings
- KernelGatewayImageConfig

- KernelSpec
- LabelCounters
- LabelCountersForWorkteam
- LabelingJobAlgorithmsConfig
- LabelingJobDataAttributes
- LabelingJobDataSource
- LabelingJobForWorkteamSummary
- LabelingJobInputConfig
- LabelingJobOutput
- LabelingJobOutputConfig
- LabelingJobResourceConfig
- LabelingJobS3DataSource
- LabelingJobSnsDataSource
- LabelingJobStoppingConditions
- LabelingJobSummary
- LambdaStepMetadata
- LastUpdateStatus
- LineageGroupSummary
- LineageMetadata
- ManagedConfiguration
- MemberDefinition
- MetadataProperties
- MetricData
- MetricDatum
- MetricDefinition
- MetricsConfig
- MetricsEndpoint
- MetricSpecification
- MetricsSource
- MlflowAppSummary

- MlflowConfig
- MLflowConfiguration
- MlflowDetails
- Model
- ModelAccessConfig
- ModelArtifacts
- ModelBiasAppSpecification
- ModelBiasBaselineConfig
- ModelBiasJobInput
- ModelCard
- ModelCardExportArtifacts
- ModelCardExportJobSummary
- ModelCardExportOutputConfig
- ModelCardSecurityConfig
- ModelCardSummary
- ModelCardVersionSummary
- ModelClientConfig
- ModelCompilationConfig
- ModelConfiguration
- ModelDashboardEndpoint
- ModelDashboardIndicatorAction
- ModelDashboardModel
- ModelDashboardModelCard
- ModelDashboardMonitoringSchedule
- ModelDataQuality
- ModelDataSource
- ModelDeployConfig
- ModelDeployResult
- ModelDigests
- ModelExplainabilityAppSpecification

- ModelExplainabilityBaselineConfig
- ModelExplainabilityJobInput
- ModelInfrastructureConfig
- ModelInput
- ModelLatencyThreshold
- ModelLifeCycle
- ModelMetadataFilter
- ModelMetadataSearchExpression
- ModelMetadataSummary
- ModelMetrics
- ModelPackage
- ModelPackageConfig
- ModelPackageContainerDefinition
- ModelPackageGroup
- ModelPackageGroupSummary
- ModelPackageModelCard
- ModelPackageSecurityConfig
- ModelPackageStatusDetails
- ModelPackageStatusItem
- ModelPackageSummary
- ModelPackageValidationProfile
- ModelPackageValidationSpecification
- ModelQuality
- ModelQualityAppSpecification
- ModelQualityBaselineConfig
- ModelQualityJobInput
- ModelQuantizationConfig
- ModelRegisterSettings
- ModelShardingConfig
- ModelSpeculativeDecodingConfig

- ModelSpeculativeDecodingTrainingDataSource
- ModelStepMetadata
- ModelSummary
- ModelVariantConfig
- ModelVariantConfigSummary
- MonitoringAlertActions
- MonitoringAlertHistorySummary
- MonitoringAlertSummary
- MonitoringAppSpecification
- MonitoringBaselineConfig
- MonitoringClusterConfig
- MonitoringConstraintsResource
- MonitoringCsvDatasetFormat
- MonitoringDatasetFormat
- MonitoringExecutionSummary
- MonitoringGroundTruthS3Input
- MonitoringInput
- MonitoringJobDefinition
- MonitoringJobDefinitionSummary
- MonitoringJsonDatasetFormat
- MonitoringNetworkConfig
- MonitoringOutput
- MonitoringOutputConfig
- MonitoringParquetDatasetFormat
- MonitoringResources
- MonitoringS3Output
- MonitoringSchedule
- MonitoringScheduleConfig
- MonitoringScheduleSummary
- MonitoringStatisticsResource

- MonitoringStoppingCondition
- MultiModelConfig
- NeoVpcConfig
- NestedFilters
- NetworkConfig
- NodeAdditionResult
- NotebookInstanceLifecycleConfigSummary
- NotebookInstanceLifecycleHook
- NotebookInstanceSummary
- NotificationConfiguration
- ObjectiveStatusCounters
- OfflineStoreConfig
- OfflineStoreStatus
- OidcConfig
- OidcConfigForResponse
- OidcMemberDefinition
- OnlineStoreConfig
- OnlineStoreConfigUpdate
- OnlineStoreSecurityConfig
- OptimizationConfig
- OptimizationJobModelSource
- OptimizationJobModelSourceS3
- OptimizationJobOutputConfig
- OptimizationJobSummary
- OptimizationModelAccessConfig
- OptimizationOutput
- OptimizationSageMakerModel
- OptimizationVpcConfig
- OutputConfig
- OutputDataConfig

- OutputParameter
- OwnershipSettings
- OwnershipSettingsSummary
- ParallelismConfiguration
- Parameter
- ParameterRange
- ParameterRanges
- Parent
- ParentHyperParameterTuningJob
- PartnerAppConfig
- PartnerAppMaintenanceConfig
- PartnerAppSummary
- PendingDeploymentSummary
- PendingProductionVariantSummary
- Phase
- Pipeline
- PipelineDefinitionS3Location
- PipelineExecution
- PipelineExecutionStep
- PipelineExecutionStepMetadata
- PipelineExecutionSummary
- PipelineExperimentConfig
- PipelineSummary
- PipelineVersion
- PipelineVersionSummary
- PlacementSpecification
- PredefinedMetricSpecification
- PresignedUrlAccessConfig
- PriorityClass
- ProcessingClusterConfig

- ProcessingFeatureStoreOutput
- ProcessingInput
- ProcessingJob
- ProcessingJobStepMetadata
- ProcessingJobSummary
- ProcessingOutput
- ProcessingOutputConfig
- ProcessingResources
- ProcessingS3Input
- ProcessingS3Output
- ProcessingStoppingCondition
- ProductionVariant
- ProductionVariantCapacityReservationConfig
- ProductionVariantCapacityReservationSummary
- ProductionVariantCoreDumpConfig
- ProductionVariantManagedInstanceScaling
- ProductionVariantManagedInstanceScalingScaleInPolicy
- ProductionVariantRoutingConfig
- ProductionVariantServerlessConfig
- ProductionVariantServerlessUpdateConfig
- ProductionVariantStatus
- ProductionVariantSummary
- ProfilerConfig
- ProfilerConfigForUpdate
- ProfilerRuleConfiguration
- ProfilerRuleEvaluationStatus
- Project
- ProjectSummary
- PropertyNameQuery
- PropertyNameSuggestion

- ProvisioningParameter
- PublicWorkforceTaskPrice
- QualityCheckStepMetadata
- QueryFilters
- RealTimeInferenceConfig
- RealTimeInferenceRecommendation
- RecommendationJobCompiledOutputConfig
- RecommendationJobContainerConfig
- RecommendationJobInferenceBenchmark
- RecommendationJobInputConfig
- RecommendationJobOutputConfig
- RecommendationJobPayloadConfig
- RecommendationJobResourceLimit
- RecommendationJobStoppingConditions
- RecommendationJobVpcConfig
- RecommendationMetrics
- RedshiftDatasetDefinition
- RegisterModelStepMetadata
- RemoteDebugConfig
- RemoteDebugConfigForUpdate
- RenderableTask
- RenderingError
- RepositoryAuthConfig
- ReservedCapacityOffering
- ReservedCapacitySummary
- ResolvedAttributes
- ResourceCatalog
- ResourceConfig
- ResourceConfigForUpdate
- ResourceLimits

- ResourceSharingConfig
- ResourceSpec
- RetentionPolicy
- RetryStrategy
- RoleGroupAssignment
- RollingDeploymentPolicy
- RollingUpdatePolicy
- RSessionAppSettings
- RStudioServerProAppSettings
- RStudioServerProDomainSettings
- RStudioServerProDomainSettingsForUpdate
- S3DataSource
- S3FileSystem
- S3FileSystemConfig
- S3ModelDataSource
- S3Presign
- S3StorageConfig
- ScalingPolicy
- ScalingPolicyMetric
- ScalingPolicyObjective
- ScheduleConfig
- ScheduledUpdateConfig
- SchedulerConfig
- SearchExpression
- SearchRecord
- SecondaryStatusTransition
- SelectedStep
- SelectiveExecutionConfig
- SelectiveExecutionResult
- ServerlessJobConfig

- ServiceCatalogProvisionedProductDetails
- ServiceCatalogProvisioningDetails
- ServiceCatalogProvisioningUpdateDetails
- SessionChainingConfig
- ShadowModeConfig
- ShadowModelVariantConfig
- SharingSettings
- ShuffleConfig
- SourceAlgorithm
- SourceAlgorithmSpecification
- SourceIpConfig
- SpaceAppLifecycleManagement
- SpaceCodeEditorAppSettings
- SpaceDetails
- SpaceIdleSettings
- SpaceJupyterLabAppSettings
- SpaceSettings
- SpaceSettingsSummary
- SpaceSharingSettings
- SpaceSharingSettingsSummary
- SpaceStorageSettings
- Stairs
- StoppingCondition
- StudioLifecycleConfigDetails
- StudioWebPortalSettings
- SubscribedWorkteam
- SuggestionQuery
- TabularJobConfig
- TabularResolvedAttributes
- Tag

- TargetPlatform
- TargetTrackingScalingPolicyConfiguration
- TemplateProviderDetail
- TensorBoardAppSettings
- TensorBoardOutputConfig
- TextClassificationJobConfig
- TextGenerationJobConfig
- TextGenerationResolvedAttributes
- ThroughputConfig
- ThroughputConfigDescription
- ThroughputConfigUpdate
- TimeSeriesConfig
- TimeSeriesForecastingJobConfig
- TimeSeriesForecastingSettings
- TimeSeriesTransformations
- TotalHits
- TrackingServerSummary
- TrafficPattern
- TrafficRoutingConfig
- TrainingImageConfig
- TrainingJob
- TrainingJobDefinition
- TrainingJobStatusCounters
- TrainingJobStepMetadata
- TrainingJobSummary
- TrainingPlanExtension
- TrainingPlanExtensionOffering
- TrainingPlanFilter
- TrainingPlanOffering
- TrainingPlanSummary

- TrainingProgressInfo
- TrainingRepositoryAuthConfig
- TrainingSpecification
- TransformDataSource
- TransformInput
- TransformJob
- TransformJobDefinition
- TransformJobStepMetadata
- TransformJobSummary
- TransformOutput
- TransformResources
- TransformS3DataSource
- Trial
- TrialComponent
- TrialComponentArtifact
- TrialComponentMetricSummary
- TrialComponentParameterValue
- TrialComponentSimpleSummary
- TrialComponentSource
- TrialComponentSourceDetail
- TrialComponentStatus
- TrialComponentSummary
- TrialSource
- TrialSummary
- TrustedIdentityPropagationSettings
- TtlDuration
- TuningJobCompletionCriteria
- TuningJobStepMetaData
- UiConfig
- UiTemplate

- UiTemplateInfo
- UltraServer
- UltraServerInfo
- UltraServerSummary
- UnifiedStudioSettings
- UpdateClusterSoftwareInstanceGroupSpecification
- UpdateTemplateProvider
- USD
- UserContext
- UserProfileDetails
- UserSettings
- VariantProperty
- VectorConfig
- Vertex
- VisibilityConditions
- VpcConfig
- WarmPoolStatus
- WorkerAccessConfiguration
- Workforce
- WorkforceVpcConfigRequest
- WorkforceVpcConfigResponse
- WorkloadSpec
- WorkspaceSettings
- Workteam

AcceleratorPartitionConfig

Service: Amazon SageMaker Service

Configuration for allocating accelerator partitions.

Contents

Count

The number of accelerator partitions to allocate with the specified partition type. If you don't specify a value for vCPU and MemoryInGiB, SageMaker AI automatically allocates ratio-based values for those parameters based on the accelerator partition count you provide.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: Yes

Type

The Multi-Instance GPU (MIG) profile type that defines the partition configuration. The profile specifies the compute and memory allocation for each partition instance. The available profile types depend on the instance type specified in the compute quota configuration.

Type: String

Valid Values: `mig-1g.5gb | mig-1g.10gb | mig-1g.18gb | mig-1g.20gb | mig-1g.23gb | mig-1g.35gb | mig-1g.45gb | mig-1g.47gb | mig-2g.10gb | mig-2g.20gb | mig-2g.35gb | mig-2g.45gb | mig-2g.47gb | mig-3g.20gb | mig-3g.40gb | mig-3g.71gb | mig-3g.90gb | mig-3g.93gb | mig-4g.20gb | mig-4g.40gb | mig-4g.71gb | mig-4g.90gb | mig-4g.93gb | mig-7g.40gb | mig-7g.80gb | mig-7g.141gb | mig-7g.180gb | mig-7g.186gb`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ActionSource

Service: Amazon SageMaker Service

A structure describing the source of an action.

Contents

SourceUri

The URI of the source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*`

Required: Yes

SourceId

The ID of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

SourceType

The type of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ActionSummary

Service: Amazon SageMaker Service

Lists the properties of an *action*. An action represents an action or activity. Some examples are a workflow step and a model deployment. Generally, an action involves at least one input artifact or output artifact.

Contents

ActionArn

The Amazon Resource Name (ARN) of the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:action/.*`

Required: No

ActionName

The name of the action.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

ActionType

The type of the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Required: No

CreationTime

When the action was created.

Type: Timestamp

Required: No

LastModifiedTime

When the action was last modified.

Type: Timestamp

Required: No

Source

The source of the action.

Type: ActionSource object

Required: No

Status

The status of the action.

Type: String

Valid Values: `Unknown | InProgress | Completed | Failed | Stopping | Stopped`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AddClusterNodeSpecification

Service: Amazon SageMaker Service

Specifies an instance group and the number of nodes to add to it.

Contents

IncrementTargetCountBy

The number of nodes to add to the specified instance group. The total number of nodes across all instance groups in a single request cannot exceed 50.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 50.

Required: Yes

InstanceGroupName

The name of the instance group to which you want to add nodes.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

AvailabilityZones

The availability zones in which to add nodes. Use this to target node placement in specific availability zones within a flexible instance group.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Pattern: `[a-z]{2}-[a-z]+-\d[a-z]`

Required: No

InstanceTypes

The instance types to use when adding nodes. Use this to target specific instance types within a flexible instance group.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge | ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge |`

`ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AdditionalEnis

Service: Amazon SageMaker Service

Information about additional Elastic Network Interfaces (ENIs) associated with an instance.

Contents

EfaEnis

A list of Elastic Fabric Adapter (EFA) ENIs associated with the instance.

Type: Array of strings

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AdditionalInferenceSpecificationDefinition

Service: Amazon SageMaker Service

A structure of additional Inference Specification. Additional Inference Specification specifies details about inference jobs that can be run with models based on this model package

Contents

Containers

The Amazon ECR registry path of the Docker image that contains the inference code.

Type: Array of ModelPackageContainerDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 15 items.

Required: Yes

Name

A unique name to identify the additional inference specification. The name must be unique within the list of your additional inference specifications for a particular model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Description

A description of the additional Inference specification

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

SupportedContentTypes

The supported MIME types for the input data.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

SupportedRealtimeInferenceInstanceTypes

A list of the instance types that are used to generate inferences in real-time.

Type: Array of strings

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge`

| ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large | ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge | ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large | ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge | ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge | ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge | ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge | ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge | ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge | ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large | ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge | ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge | ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge | ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge | ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium | ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge | ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge | ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge | ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge

| ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge | ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge | ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large | ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge | ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge | ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge

Required: No

SupportedResponseMIMETypes

The supported MIME types for the output data.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[-\w]+\/.+`

Required: No

SupportedTransformInstanceTypes

A list of the instance types on which a transformation job can be run or on which an endpoint can be deployed.

Type: Array of strings

Array Members: Minimum number of 1 item.

Valid Values: ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge

```
| ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge |
ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge
| ml.r6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge |
ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge
| ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge
| ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge
| ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large
| ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge |
ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge
| ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge
| ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.g5.xlarge | ml.g5.2xlarge
| ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge |
ml.g5.24xlarge | ml.g5.48xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge |
ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge
| ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge |
ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge
```

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AdditionalModelDataSource

Service: Amazon SageMaker Service

Data sources that are available to your model in addition to the one that you specify for `ModelDataSource` when you use the `CreateModel` action.

Contents

ChannelName

A custom name for this `AdditionalModelDataSource` object.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[A-Za-z0-9\.\-_]+`

Required: Yes

S3DataSource

Specifies the S3 location of ML model data to deploy.

Type: S3ModelDataSource object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AdditionalS3DataSource

Service: Amazon SageMaker Service

A data source used for training or inference that is in addition to the input dataset or model data.

Contents

S3DataType

The data type of the additional data source that you specify for use in inference or training.

Type: String

Valid Values: `S3Object | S3Prefix`

Required: Yes

S3Uri

The uniform resource identifier (URI) used to identify an additional data source used in inference or training.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

CompressionType

The type of compression used for an additional data source used in inference or training. Specify `None` if your additional data source is not compressed.

Type: String

Valid Values: `None | Gzip`

Required: No

ETag

The ETag associated with S3 URI.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentVersion

Service: Amazon SageMaker Service

Edge Manager agent version.

Contents

AgentCount

The number of Edge Manager agents.

Type: Long

Required: Yes

Version

Version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 30.

Pattern: `[a-zA-Z0-9\ \_\.]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIBenchmarkEndpoint

Service: Amazon SageMaker Service

The SageMaker endpoint configuration for benchmarking.

Contents

Identifier

The name or Amazon Resource Name (ARN) of the SageMaker endpoint to benchmark.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*/)?([a-zA-Z0-9]([a-zA-Z0-9\-]){0,62})(?<!-)`

Required: Yes

InferenceComponents

The list of inference components to benchmark on the endpoint.

Type: Array of AIBenchmarkInferenceComponent objects

Required: No

TargetContainerHostname

The hostname of the specific container to target within a multi-container endpoint.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AIBenchmarkInferenceComponent

Service: Amazon SageMaker Service

An inference component to benchmark.

Contents

Identifier

The name or Amazon Resource Name (ARN) of the inference component.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*/)?([a-zA-Z0-9]([a-zA-Z0-9\-]){0,62})(?<!-)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIBenchmarkJobSummary

Service: Amazon SageMaker Service

Summary information about an AI benchmark job.

Contents

AIBenchmarkJobArn

The Amazon Resource Name (ARN) of the benchmark job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-benchmark-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AIBenchmarkJobName

The name of the benchmark job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AIBenchmarkJobStatus

The status of the benchmark job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: Yes

CreationTime

A timestamp that indicates when the benchmark job was created.

Type: Timestamp

Required: Yes

AIWorkloadConfigName

The name of the AI workload configuration used by the benchmark job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

EndTime

A timestamp that indicates when the benchmark job completed.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIBenchmarkNetworkConfig

Service: Amazon SageMaker Service

The network configuration for an AI benchmark job.

Contents

VpcConfig

The VPC configuration, including security group IDs and subnet IDs.

Type: VpcConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIBenchmarkOutputConfig

Service: Amazon SageMaker Service

The output configuration for an AI benchmark job.

Contents

S3OutputLocation

The Amazon S3 URI where benchmark results are stored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

MlflowConfig

The MLflow tracking configuration for the job. If you don't specify this parameter, MLflow tracking is disabled.

Type: AIMlflowConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIBenchmarkOutputResult

Service: Amazon SageMaker Service

The output result of an AI benchmark job, including the Amazon S3 location and CloudWatch log information.

Contents

S3OutputLocation

The Amazon S3 URI where benchmark results are stored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

CloudWatchLogs

The CloudWatch log information for the benchmark job.

Type: Array of AICloudWatchLogs objects

Required: No

MlflowConfig

The MLflow tracking configuration for the job.

Type: AIMlflowConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AIBenchmarkTarget

Service: Amazon SageMaker Service

The target for an AI benchmark job. This is a union type — specify one of the members.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

Endpoint

The SageMaker endpoint to benchmark.

Type: AIBenchmarkEndpoint object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AICapacityReservationConfig

Service: Amazon SageMaker Service

The capacity reservation configuration for an AI recommendation job.

Contents

CapacityReservationPreference

The capacity reservation preference. The only valid value is `capacity-reservations-only`.

Type: String

Valid Values: `capacity-reservations-only`

Required: No

MlReservationArns

The list of ML reservation ARNs to use.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z0-9\-]{1,14}/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AICloudWatchLogs

Service: Amazon SageMaker Service

CloudWatch log information for an AI benchmark or recommendation job.

Contents

LogGroupArn

The Amazon Resource Name (ARN) of the CloudWatch log group.

Type: String

Required: No

LogStreamName

The name of the CloudWatch log stream.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIDatasetConfig

Service: Amazon SageMaker Service

The dataset configuration for an AI workload. This is a union type — specify one of the members.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

InputDataConfig

An array of input data channel configurations for the workload.

Type: Array of AIWorkloadInputDataConfig objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIMlflowConfig

Service: Amazon SageMaker Service

The MLflow tracking configuration for logging metrics and parameters to a SageMaker managed MLflow tracking server.

Contents

MlflowResourceArn

The Amazon Resource Name (ARN) of the SageMaker managed MLflow resource.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-(app|tracking-server)/.*`

Required: Yes

MlflowExperimentName

The MLflow experiment name used for tracking.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9\-_./]+`

Required: No

MlflowRunName

The MLflow run name used for tracking.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9\-_./]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIModelSource

Service: Amazon SageMaker Service

The source of the model for an AI recommendation job. This is a union type.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

S3

The Amazon S3 location of the model artifacts.

Type: AIModelSourceS3 object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIModelSourceS3

Service: Amazon SageMaker Service

The Amazon S3 model source configuration.

Contents

S3Uri

The Amazon S3 URI of the model artifacts.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendation

Service: Amazon SageMaker Service

An optimization recommendation generated by an AI recommendation job.

Contents

AIBenchmarkJobArn

The Amazon Resource Name (ARN) of the benchmark job associated with this recommendation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-benchmark-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

DeploymentConfiguration

The deployment configuration for this recommendation, including the container image, instance type, instance count, and environment variables.

Type: AIRecommendationDeploymentConfiguration object

Required: No

ExpectedPerformance

The expected performance metrics for this recommendation.

Type: Array of AIRecommendationPerformanceMetric objects

Required: No

ModelDetails

Details about the model package associated with this recommendation.

Type: AIRecommendationModelDetails object

Required: No

OptimizationDetails

The optimization techniques applied in this recommendation.

Type: Array of AIRecommendationOptimizationDetail objects

Required: No

RecommendationDescription

A description of the recommendation.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationComputeSpec

Service: Amazon SageMaker Service

The compute resource specification for an AI recommendation job.

Contents

CapacityReservationConfig

The capacity reservation configuration.

Type: AICapacityReservationConfig object

Required: No

InstanceTypes

The list of instance types to consider for recommendations. You can specify up to 3 instance types.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 3 items.

Valid Values: `ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.4xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationConstraint

Service: Amazon SageMaker Service

A performance constraint for an AI recommendation job.

Contents

Metric

The performance metric. Valid values are `ttft-ms` (time to first token in milliseconds), `throughput`, and `cost`.

Type: String

Valid Values: `ttft-ms | throughput | cost`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationDeploymentConfiguration

Service: Amazon SageMaker Service

The deployment configuration for a recommendation.

Contents

CopyCountPerInstance

The number of model copies per instance.

Type: Integer

Required: No

EnvironmentVariables

The environment variables for the deployment.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 1024.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 1024.

Value Pattern: `[\S\s]*`

Required: No

ImageUri

The URI of the container image for the deployment.

Type: String

Required: No

InstanceCount

The recommended number of instances for the deployment.

Type: Integer

Required: No

InstanceType

The recommended instance type for the deployment.

Type: String

Valid Values: `ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.4xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge`

Required: No

S3

The Amazon S3 data channels for the deployment.

Type: Array of AIRecommendationDeploymentS3Channel objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationDeploymentS3Channel

Service: Amazon SageMaker Service

An Amazon S3 data channel for a recommended deployment configuration, containing model artifacts or optimized model outputs.

Contents

ChannelName

A custom name for this Amazon S3 data channel.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[A-Za-z0-9\.\-_]+`

Required: No

Uri

The Amazon S3 URI of the data for this channel.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationInferenceSpecification

Service: Amazon SageMaker Service

The inference framework for an AI recommendation job.

Contents

Framework

The inference framework. Valid values are `LMI` and `VLLM`.

Type: String

Valid Values: `LMI | VLLM`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationInstanceDetail

Service: Amazon SageMaker Service

Instance details for a recommendation.

Contents

CopyCountPerInstance

The number of model copies per instance.

Type: Integer

Required: No

InstanceCount

The recommended number of instances.

Type: Integer

Required: No

InstanceType

The recommended instance type.

Type: String

Valid Values: `ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.4xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationJobSummary

Service: Amazon SageMaker Service

Summary information about an AI recommendation job.

Contents

AIRecommendationJobArn

The Amazon Resource Name (ARN) of the recommendation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-recommendation-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AIRecommendationJobName

The name of the recommendation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AIRecommendationJobStatus

The status of the recommendation job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: Yes

CreationTime

A timestamp that indicates when the recommendation job was created.

Type: Timestamp

Required: Yes

EndTime

A timestamp that indicates when the recommendation job completed.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationModelDetails

Service: Amazon SageMaker Service

Details about the model package in a recommendation.

Contents

InferenceSpecificationName

The name of the inference specification within the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

InstanceDetails

The instance details for this recommendation, including instance type, count, and model copies per instance.

Type: Array of AIRecommendationInstanceDetail objects

Required: No

ModelPackageArn

The Amazon Resource Name (ARN) of the model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationOptimizationDetail

Service: Amazon SageMaker Service

Details about an optimization technique applied in a recommendation.

Contents

OptimizationType

The type of optimization. Valid values are `SpeculativeDecoding` and `KernelTuning`.

Type: String

Valid Values: `SpeculativeDecoding | KernelTuning`

Required: Yes

OptimizationConfig

A map of configuration parameters for the optimization technique.

Type: String to string map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationOutputConfig

Service: Amazon SageMaker Service

The output configuration for an AI recommendation job.

Contents

MlflowConfig

The MLflow tracking configuration for the job. If you don't specify this parameter, MLflow tracking is disabled.

Type: AIMlflowConfig object

Required: No

ModelPackageGroupIdentifier

The name or Amazon Resource Name (ARN) of the model package group where the optimized model is registered as a new model package version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*/)?([a-zA-Z0-9]([a-zA-Z0-9\-]){0,62})(?<!-)`

Required: No

S3OutputLocation

The Amazon S3 URI where recommendation results are stored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationOutputResult

Service: Amazon SageMaker Service

The output configuration for an AI recommendation job, including the S3 location for results and the model package group for deployment.

Contents

S3OutputLocation

The Amazon S3 URI where the recommendation job writes its output results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

MlflowConfig

The MLflow tracking configuration for the job.

Type: AIMlflowConfig object

Required: No

ModelPackageGroupIdentifier

The name or Amazon Resource Name (ARN) of the model package group where deployment-ready model packages are registered.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*/)?([a-zA-Z0-9]([a-zA-Z0-9\-]){0,62})(?<!-)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationPerformanceMetric

Service: Amazon SageMaker Service

An expected performance metric for a recommendation.

Contents

Metric

The name of the performance metric.

Type: String

Required: Yes

Value

The value of the metric.

Type: String

Required: Yes

Stat

The statistical measure for the metric.

Type: String

Required: No

Unit

The unit of the metric value.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIRecommendationPerformanceTarget

Service: Amazon SageMaker Service

The performance targets for an AI recommendation job.

Contents

Constraints

An array of performance constraints that define the optimization objectives.

Type: Array of AIRecommendationConstraint objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIWorkloadConfigs

Service: Amazon SageMaker Service

The benchmark tool configuration for an AI workload.

Contents

WorkloadSpec

The workload specification that defines benchmark parameters.

Type: WorkloadSpec object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIWorkloadConfigSummary

Service: Amazon SageMaker Service

Summary information about an AI workload configuration.

Contents

AIWorkloadConfigArn

The Amazon Resource Name (ARN) of the AI workload configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:ai-workload-config/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AIWorkloadConfigName

The name of the AI workload configuration.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

CreationTime

A timestamp that indicates when the configuration was created.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIWorkloadDataSource

Service: Amazon SageMaker Service

The data source for an AI workload input data channel.

Contents

S3DataSource

The Amazon S3 data source configuration.

Type: AIWorkloadS3DataSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIWorkloadInputDataConfig

Service: Amazon SageMaker Service

A channel of input data for an AI workload configuration. Each channel has a name and a data source.

Contents

ChannelName

The logical name for the data channel.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[A-Za-z0-9\.\-_]+`

Required: Yes

DataSource

The data source for this channel.

Type: AIWorkloadDataSource object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AIWorkloadS3DataSource

Service: Amazon SageMaker Service

The Amazon S3 data source for an AI workload.

Contents

S3Uri

The Amazon S3 URI of the data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Alarm

Service: Amazon SageMaker Service

An Amazon CloudWatch alarm configured to monitor metrics on an endpoint.

Contents

AlarmName

The name of a CloudWatch alarm in your account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `(?!\s*$).+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AlarmDetails

Service: Amazon SageMaker Service

The details of the alarm to monitor during the AMI update.

Contents

AlarmName

The name of the alarm.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `(?!\s*$).+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AlgorithmSpecification

Service: Amazon SageMaker Service

Specifies the training algorithm to use in a CreateTrainingJob request.

> **Important**
>
> SageMaker uses its own SageMaker account credentials to pull and access built-in algorithms so built-in algorithms are universally accessible across all AWS accounts. As a result, built-in algorithms have standard, unrestricted access. You cannot restrict built-in algorithms using IAM roles. Use custom algorithms if you require specific access controls.

For more information about algorithms provided by SageMaker, see Algorithms. For information about using your own algorithms, see Using Your Own Algorithms with Amazon SageMaker.

Contents

TrainingInputMode

The training input mode that the algorithm supports. For more information about input modes, see Algorithms.

Pipe mode

If an algorithm supports `Pipe` mode, Amazon SageMaker streams data directly from Amazon S3 to the container.

File mode

If an algorithm supports `File` mode, SageMaker downloads the training data from S3 to the provisioned ML storage volume, and mounts the directory to the Docker volume for the training container.

You must provision the ML storage volume with sufficient capacity to accommodate the data downloaded from S3. In addition to the training data, the ML storage volume also stores the output model. The algorithm container uses the ML storage volume to also store intermediate information, if any.

For distributed algorithms, training data is distributed uniformly. Your training duration is predictable if the input data objects sizes are approximately the same. SageMaker does not split

the files any further for model training. If the object sizes are skewed, training won't be optimal as the data distribution is also skewed when one host in a training cluster is overloaded, thus becoming a bottleneck in training.

FastFile mode

If an algorithm supports `FastFile` mode, SageMaker streams data directly from S3 to the container with no code changes, and provides file system access to the data. Users can author their training script to interact with these files as if they were stored on disk.

`FastFile` mode works best when the data is read sequentially. Augmented manifest files aren't supported. The startup time is lower when there are fewer files in the S3 bucket provided.

Type: String

Valid Values: `Pipe | File | FastFile`

Required: Yes

AlgorithmName

The name of the algorithm resource to use for the training job. This must be an algorithm resource that you created or subscribe to on AWS Marketplace.

> **Note**
>
> You must specify either the algorithm name to the `AlgorithmName` parameter or the image URI of the algorithm container to the `TrainingImage` parameter.
> Note that the `AlgorithmName` parameter is mutually exclusive with the `TrainingImage` parameter. If you specify a value for the `AlgorithmName` parameter, you can't specify a value for `TrainingImage`, and vice versa.
> If you specify values for both parameters, the training job might break; if you don't specify any value for both parameters, the training job might raise a `null` error.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 170.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*\/)?([a-zA-Z0-9]([a-zA-Z0-9-]){0,62})(?<!-)`

Required: No

ContainerArguments

The arguments for a container used to run a training job. See How Amazon SageMaker Runs Your Training Image for additional information.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

ContainerEntrypoint

The entrypoint script for a Docker container used to run a training job. This script takes precedence over the default train processing instructions. See How Amazon SageMaker Runs Your Training Image for more information.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

EnableSageMakerMetricsTimeSeries

To generate and save time-series metrics during training, set to `true`. The default is `false` and time-series metrics aren't generated except in the following cases:

- You use one of the SageMaker built-in algorithms
- You use one of the following Prebuilt SageMaker Docker Images:
  - Tensorflow (version >= 1.15)
  - MXNet (version >= 1.6)
  - PyTorch (version >= 1.3)

- You specify at least one MetricDefinition

Type: Boolean

Required: No

MetricDefinitions

A list of metric definition objects. Each object specifies the metric name and regular expressions used to parse algorithm logs. SageMaker publishes each metric to Amazon CloudWatch.

Type: Array of MetricDefinition objects

Array Members: Minimum number of 0 items. Maximum number of 40 items.

Required: No

TrainingImage

The registry path of the Docker image that contains the training algorithm. For information about docker registry paths for SageMaker built-in algorithms, see Docker Registry Paths and Example Code in the *Amazon SageMaker developer guide*. SageMaker supports both `registry/repository[:tag]` and `registry/repository[@digest]` image path formats. For more information about using your custom training container, see Using Your Own Algorithms with Amazon SageMaker.

> **Note**
>
> You must specify either the algorithm name to the `AlgorithmName` parameter or the image URI of the algorithm container to the `TrainingImage` parameter.
> For more information, see the note in the `AlgorithmName` parameter description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

TrainingImageConfig

The configuration to use an image from a private Docker registry for a training job.

Type: TrainingImageConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AlgorithmStatusDetails

Service: Amazon SageMaker Service

Specifies the validation and image scan statuses of the algorithm.

Contents

ImageScanStatuses

The status of the scan of the algorithm's Docker image container.

Type: Array of AlgorithmStatusItem objects

Required: No

ValidationStatuses

The status of algorithm validation.

Type: Array of AlgorithmStatusItem objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AlgorithmStatusItem

Service: Amazon SageMaker Service

Represents the overall status of an algorithm.

Contents

Name

The name of the algorithm for which the overall status is being reported.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Status

The current status.

Type: String

Valid Values: `NotStarted | InProgress | Completed | Failed`

Required: Yes

FailureReason

if the overall status is `Failed`, the reason for the failure.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

AlgorithmSummary

Service: Amazon SageMaker Service

Provides summary information about an algorithm.

Contents

AlgorithmArn

The Amazon Resource Name (ARN) of the algorithm.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:algorithm/[\S]{1,2048}`

Required: Yes

AlgorithmName

The name of the algorithm that is described by the summary.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AlgorithmStatus

The overall status of the algorithm.

Type: String

Valid Values: `Pending | InProgress | Completed | Failed | Deleting`

Required: Yes

CreationTime

A timestamp that shows when the algorithm was created.

Type: Timestamp

Required: Yes

AlgorithmDescription

A brief description of the algorithm.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AlgorithmValidationProfile

Service: Amazon SageMaker Service

Defines a training job and a batch transform job that SageMaker runs to validate your algorithm.

The data provided in the validation profile is made available to your buyers on AWS Marketplace.

Contents

ProfileName

The name of the profile for the algorithm. The name must have 1 to 63 characters. Valid characters are a-z, A-Z, 0-9, and - (hyphen).

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

TrainingJobDefinition

The `TrainingJobDefinition` object that describes the training job that SageMaker runs to validate your algorithm.

Type: TrainingJobDefinition object

Required: Yes

TransformJobDefinition

The `TransformJobDefinition` object that describes the transform job that SageMaker runs to validate your algorithm.

Type: TransformJobDefinition object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AlgorithmValidationSpecification

Service: Amazon SageMaker Service

Specifies configurations for one or more training jobs that SageMaker runs to test the algorithm.

Contents

ValidationProfiles

An array of `AlgorithmValidationProfile` objects, each of which specifies a training job and batch transform job that SageMaker runs to validate your algorithm.

Type: Array of AlgorithmValidationProfile objects

Array Members: Fixed number of 1 item.

Required: Yes

ValidationRole

The IAM roles that SageMaker uses to run the training jobs.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AmazonQSettings

Service: Amazon SageMaker Service

A collection of settings that configure the Amazon Q experience within the domain.

Contents

QProfileArn

The ARN of the Amazon Q profile used within the domain.

Type: String

Pattern: `arn:[-.a-z0-9]{1,63}:codewhisperer:([-.a-z0-9]{0,63}:){2}([a-zA-Z0-9-_:/]){1,1023}`

Required: No

Status

Whether Amazon Q has been enabled within the domain.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AnnotationConsolidationConfig

Service: Amazon SageMaker Service

Configures how labels are consolidated across human workers and processes output data.

Contents

AnnotationConsolidationLambdaArn

The Amazon Resource Name (ARN) of a Lambda function implements the logic for annotation consolidation and to process output data.

For built-in task types, use one of the following Amazon SageMaker Ground Truth Lambda function ARNs for `AnnotationConsolidationLambdaArn`. For custom labeling workflows, see Post-annotation Lambda.

Bounding box - Finds the most similar boxes from different workers based on the Jaccard index of the boxes.

- `arn:aws:lambda:us-east-1:432418664414:function:ACS-BoundingBox`
- `arn:aws:lambda:us-east-2:266458841044:function:ACS-BoundingBox`
- `arn:aws:lambda:us-west-2:081040173940:function:ACS-BoundingBox`
- `arn:aws:lambda:eu-west-1:568282634449:function:ACS-BoundingBox`
- `arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-BoundingBox`
- `arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-BoundingBox`
- `arn:aws:lambda:ap-south-1:565803892007:function:ACS-BoundingBox`
- `arn:aws:lambda:eu-central-1:203001061592:function:ACS-BoundingBox`
- `arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-BoundingBox`
- `arn:aws:lambda:eu-west-2:487402164563:function:ACS-BoundingBox`
- `arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-BoundingBox`
- `arn:aws:lambda:ca-central-1:918755190332:function:ACS-BoundingBox`

Image classification - Uses a variant of the Expectation Maximization approach to estimate the true class of an image based on annotations from individual workers.

- `arn:aws:lambda:us-east-1:432418664414:function:ACS-ImageMultiClass`

- arn:aws:lambda:us-east-2:266458841044:function:ACS-ImageMultiClass
- arn:aws:lambda:us-west-2:081040173940:function:ACS-ImageMultiClass
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-ImageMultiClass
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-ImageMultiClass
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-ImageMultiClass
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-ImageMultiClass
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-ImageMultiClass
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-ImageMultiClass
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-ImageMultiClass
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-ImageMultiClass
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-ImageMultiClass

Multi-label image classification - Uses a variant of the Expectation Maximization approach to estimate the true classes of an image based on annotations from individual workers.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-ImageMultiClassMultiLabel
- arn:aws:lambda:us-east-2:266458841044:function:ACS-ImageMultiClassMultiLabel
- arn:aws:lambda:us-west-2:081040173940:function:ACS-ImageMultiClassMultiLabel
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-ImageMultiClassMultiLabel
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-ImageMultiClassMultiLabel
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-ImageMultiClassMultiLabel
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-ImageMultiClassMultiLabel

- arn:aws:lambda:eu-central-1:203001061592:function:ACS-ImageMultiClassMultiLabel
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-ImageMultiClassMultiLabel
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-ImageMultiClassMultiLabel
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-ImageMultiClassMultiLabel
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-ImageMultiClassMultiLabel

Semantic segmentation - Treats each pixel in an image as a multi-class classification and treats pixel annotations from workers as "votes" for the correct label.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-SemanticSegmentation
- arn:aws:lambda:us-east-2:266458841044:function:ACS-SemanticSegmentation
- arn:aws:lambda:us-west-2:081040173940:function:ACS-SemanticSegmentation
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-SemanticSegmentation
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-SemanticSegmentation
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-SemanticSegmentation
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-SemanticSegmentation
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-SemanticSegmentation
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-SemanticSegmentation
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-SemanticSegmentation

- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-SemanticSegmentation
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-SemanticSegmentation

Text classification - Uses a variant of the Expectation Maximization approach to estimate the true class of text based on annotations from individual workers.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-TextMultiClass
- arn:aws:lambda:us-east-2:266458841044:function:ACS-TextMultiClass
- arn:aws:lambda:us-west-2:081040173940:function:ACS-TextMultiClass
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-TextMultiClass
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-TextMultiClass
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-TextMultiClass
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-TextMultiClass
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-TextMultiClass
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-TextMultiClass
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-TextMultiClass
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-TextMultiClass
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-TextMultiClass

Multi-label text classification - Uses a variant of the Expectation Maximization approach to estimate the true classes of text based on annotations from individual workers.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-TextMultiClassMultiLabel
- arn:aws:lambda:us-east-2:266458841044:function:ACS-TextMultiClassMultiLabel
- arn:aws:lambda:us-west-2:081040173940:function:ACS-TextMultiClassMultiLabel
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-TextMultiClassMultiLabel
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-TextMultiClassMultiLabel

- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-TextMultiClassMultiLabel
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-TextMultiClassMultiLabel
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-TextMultiClassMultiLabel
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-TextMultiClassMultiLabel
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-TextMultiClassMultiLabel
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-TextMultiClassMultiLabel
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-TextMultiClassMultiLabel

Named entity recognition - Groups similar selections and calculates aggregate boundaries, resolving to most-assigned label.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-NamedEntityRecognition
- arn:aws:lambda:us-east-2:266458841044:function:ACS-NamedEntityRecognition
- arn:aws:lambda:us-west-2:081040173940:function:ACS-NamedEntityRecognition
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-NamedEntityRecognition
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-NamedEntityRecognition
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-NamedEntityRecognition
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-NamedEntityRecognition
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-NamedEntityRecognition

- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-NamedEntityRecognition
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-NamedEntityRecognition
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-NamedEntityRecognition
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-NamedEntityRecognition

Video Classification - Use this task type when you need workers to classify videos using predefined labels that you specify. Workers are shown videos and are asked to choose one label for each video.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-VideoMultiClass
- arn:aws:lambda:us-east-2:266458841044:function:ACS-VideoMultiClass
- arn:aws:lambda:us-west-2:081040173940:function:ACS-VideoMultiClass
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-VideoMultiClass
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-VideoMultiClass
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-VideoMultiClass
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-VideoMultiClass
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-VideoMultiClass
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-VideoMultiClass
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-VideoMultiClass
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-VideoMultiClass
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-VideoMultiClass

Video Frame Object Detection - Use this task type to have workers identify and locate objects in a sequence of video frames (images extracted from a video) using bounding boxes. For example, you can use this task to ask workers to identify and localize various objects in a series of video frames, such as cars, bikes, and pedestrians.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-VideoObjectDetection
- arn:aws:lambda:us-east-2:266458841044:function:ACS-VideoObjectDetection
- arn:aws:lambda:us-west-2:081040173940:function:ACS-VideoObjectDetection
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-VideoObjectDetection
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-VideoObjectDetection
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-VideoObjectDetection
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-VideoObjectDetection
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-VideoObjectDetection
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-VideoObjectDetection
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-VideoObjectDetection
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-VideoObjectDetection
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-VideoObjectDetection

Video Frame Object Tracking - Use this task type to have workers track the movement of objects in a sequence of video frames (images extracted from a video) using bounding boxes. For example, you can use this task to ask workers to track the movement of objects, such as cars, bikes, and pedestrians.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-VideoObjectTracking
- arn:aws:lambda:us-east-2:266458841044:function:ACS-VideoObjectTracking
- arn:aws:lambda:us-west-2:081040173940:function:ACS-VideoObjectTracking
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-VideoObjectTracking

- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-VideoObjectTracking
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-VideoObjectTracking
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-VideoObjectTracking
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-VideoObjectTracking
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-VideoObjectTracking
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-VideoObjectTracking
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-VideoObjectTracking
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-VideoObjectTracking

3D Point Cloud Object Detection - Use this task type when you want workers to classify objects in a 3D point cloud by drawing 3D cuboids around objects. For example, you can use this task type to ask workers to identify different types of objects in a point cloud, such as cars, bikes, and pedestrians.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-3DPointCloudObjectDetection
- arn:aws:lambda:us-east-2:266458841044:function:ACS-3DPointCloudObjectDetection
- arn:aws:lambda:us-west-2:081040173940:function:ACS-3DPointCloudObjectDetection
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-3DPointCloudObjectDetection
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-3DPointCloudObjectDetection
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-3DPointCloudObjectDetection
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-3DPointCloudObjectDetection

- arn:aws:lambda:eu-central-1:203001061592:function:ACS-3DPointCloudObjectDetection
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-3DPointCloudObjectDetection
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-3DPointCloudObjectDetection
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-3DPointCloudObjectDetection
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-3DPointCloudObjectDetection

3D Point Cloud Object Tracking - Use this task type when you want workers to draw 3D cuboids around objects that appear in a sequence of 3D point cloud frames. For example, you can use this task type to ask workers to track the movement of vehicles across multiple point cloud frames.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-3DPointCloudObjectTracking
- arn:aws:lambda:us-east-2:266458841044:function:ACS-3DPointCloudObjectTracking
- arn:aws:lambda:us-west-2:081040173940:function:ACS-3DPointCloudObjectTracking
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-3DPointCloudObjectTracking
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-3DPointCloudObjectTracking
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-3DPointCloudObjectTracking
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-3DPointCloudObjectTracking
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-3DPointCloudObjectTracking
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-3DPointCloudObjectTracking

- arn:aws:lambda:eu-west-2:487402164563:function:ACS-3DPointCloudObjectTracking
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-3DPointCloudObjectTracking
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-3DPointCloudObjectTracking

3D Point Cloud Semantic Segmentation - Use this task type when you want workers to create a point-level semantic segmentation masks by painting objects in a 3D point cloud using different colors where each color is assigned to one of the classes you specify.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:us-east-2:266458841044:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:us-west-2:081040173940:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-3DPointCloudSemanticSegmentation

Use the following ARNs for Label Verification and Adjustment Jobs

Use label verification and adjustment jobs to review and adjust labels. To learn more, see Verify and Adjust Labels .

Semantic Segmentation Adjustment - Treats each pixel in an image as a multi-class classification and treats pixel adjusted annotations from workers as "votes" for the correct label.

- `arn:aws:lambda:us-east-1:432418664414:function:ACS-AdjustmentSemanticSegmentation`
- `arn:aws:lambda:us-east-2:266458841044:function:ACS-AdjustmentSemanticSegmentation`
- `arn:aws:lambda:us-west-2:081040173940:function:ACS-AdjustmentSemanticSegmentation`
- `arn:aws:lambda:eu-west-1:568282634449:function:ACS-AdjustmentSemanticSegmentation`
- `arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-AdjustmentSemanticSegmentation`
- `arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-AdjustmentSemanticSegmentation`
- `arn:aws:lambda:ap-south-1:565803892007:function:ACS-AdjustmentSemanticSegmentation`
- `arn:aws:lambda:eu-central-1:203001061592:function:ACS-AdjustmentSemanticSegmentation`
- `arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-AdjustmentSemanticSegmentation`
- `arn:aws:lambda:eu-west-2:487402164563:function:ACS-AdjustmentSemanticSegmentation`
- `arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-AdjustmentSemanticSegmentation`
- `arn:aws:lambda:ca-central-1:918755190332:function:ACS-AdjustmentSemanticSegmentation`

Semantic Segmentation Verification - Uses a variant of the Expectation Maximization approach to estimate the true class of verification judgment for semantic segmentation labels based on annotations from individual workers.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-VerificationSemanticSegmentation
- arn:aws:lambda:us-east-2:266458841044:function:ACS-VerificationSemanticSegmentation
- arn:aws:lambda:us-west-2:081040173940:function:ACS-VerificationSemanticSegmentation
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-VerificationSemanticSegmentation
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-VerificationSemanticSegmentation
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-VerificationSemanticSegmentation
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-VerificationSemanticSegmentation
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-VerificationSemanticSegmentation
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-VerificationSemanticSegmentation
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-VerificationSemanticSegmentation
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-VerificationSemanticSegmentation
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-VerificationSemanticSegmentation

Bounding Box Adjustment - Finds the most similar boxes from different workers based on the Jaccard index of the adjusted annotations.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-AdjustmentBoundingBox
- arn:aws:lambda:us-east-2:266458841044:function:ACS-AdjustmentBoundingBox
- arn:aws:lambda:us-west-2:081040173940:function:ACS-AdjustmentBoundingBox

- arn:aws:lambda:eu-west-1:568282634449:function:ACS-AdjustmentBoundingBox
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-AdjustmentBoundingBox
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-AdjustmentBoundingBox
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-AdjustmentBoundingBox
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-AdjustmentBoundingBox
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-AdjustmentBoundingBox
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-AdjustmentBoundingBox
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-AdjustmentBoundingBox
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-AdjustmentBoundingBox

Bounding Box Verification - Uses a variant of the Expectation Maximization approach to estimate the true class of verification judgement for bounding box labels based on annotations from individual workers.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-VerificationBoundingBox
- arn:aws:lambda:us-east-2:266458841044:function:ACS-VerificationBoundingBox
- arn:aws:lambda:us-west-2:081040173940:function:ACS-VerificationBoundingBox
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-VerificationBoundingBox
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-VerificationBoundingBox
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-VerificationBoundingBox

- arn:aws:lambda:ap-south-1:565803892007:function:ACS-VerificationBoundingBox
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-VerificationBoundingBox
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-VerificationBoundingBox
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-VerificationBoundingBox
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-VerificationBoundingBox
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-VerificationBoundingBox

Video Frame Object Detection Adjustment - Use this task type when you want workers to adjust bounding boxes that workers have added to video frames to classify and localize objects in a sequence of video frames.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-AdjustmentVideoObjectDetection
- arn:aws:lambda:us-east-2:266458841044:function:ACS-AdjustmentVideoObjectDetection
- arn:aws:lambda:us-west-2:081040173940:function:ACS-AdjustmentVideoObjectDetection
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-AdjustmentVideoObjectDetection
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-AdjustmentVideoObjectDetection
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-AdjustmentVideoObjectDetection
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-AdjustmentVideoObjectDetection
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-AdjustmentVideoObjectDetection
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-AdjustmentVideoObjectDetection

- arn:aws:lambda:eu-west-2:487402164563:function:ACS-AdjustmentVideoObjectDetection
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-AdjustmentVideoObjectDetection
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-AdjustmentVideoObjectDetection

Video Frame Object Tracking Adjustment - Use this task type when you want workers to adjust bounding boxes that workers have added to video frames to track object movement across a sequence of video frames.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-AdjustmentVideoObjectTracking
- arn:aws:lambda:us-east-2:266458841044:function:ACS-AdjustmentVideoObjectTracking
- arn:aws:lambda:us-west-2:081040173940:function:ACS-AdjustmentVideoObjectTracking
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-AdjustmentVideoObjectTracking
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-AdjustmentVideoObjectTracking
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-AdjustmentVideoObjectTracking
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-AdjustmentVideoObjectTracking
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-AdjustmentVideoObjectTracking
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-AdjustmentVideoObjectTracking
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-AdjustmentVideoObjectTracking
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-AdjustmentVideoObjectTracking
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-AdjustmentVideoObjectTracking

3D Point Cloud Object Detection Adjustment - Use this task type when you want workers to adjust 3D cuboids around objects in a 3D point cloud.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:us-east-2:266458841044:function:ACS-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:us-west-2:081040173940:function:ACS-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-Adjustment3DPointCloudObjectDetection

3D Point Cloud Object Tracking Adjustment - Use this task type when you want workers to adjust 3D cuboids around objects that appear in a sequence of 3D point cloud frames.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:us-east-2:266458841044:function:ACS-Adjustment3DPointCloudObjectTracking

- arn:aws:lambda:us-west-2:081040173940:function:ACS-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-Adjustment3DPointCloudObjectTracking

3D Point Cloud Semantic Segmentation Adjustment - Use this task type when you want workers to adjust a point-level semantic segmentation masks using a paint tool.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-3DPointCloudSemanticSegmentation
- arn:aws:lambda:us-east-1:432418664414:function:ACS-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:us-east-2:266458841044:function:ACS-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:us-west-2:081040173940:function:ACS-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-Adjustment3DPointCloudSemanticSegmentation

- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-Adjustment3DPointCloudSemanticSegmentation

Generative AI/Custom - Direct passthrough of output data without any transformation.

- arn:aws:lambda:us-east-1:432418664414:function:ACS-PassThrough
- arn:aws:lambda:us-east-2:266458841044:function:ACS-PassThrough
- arn:aws:lambda:us-west-2:081040173940:function:ACS-PassThrough
- arn:aws:lambda:eu-west-1:568282634449:function:ACS-PassThrough
- arn:aws:lambda:ap-northeast-1:477331159723:function:ACS-PassThrough
- arn:aws:lambda:ap-southeast-2:454466003867:function:ACS-PassThrough
- arn:aws:lambda:ap-south-1:565803892007:function:ACS-PassThrough
- arn:aws:lambda:eu-central-1:203001061592:function:ACS-PassThrough
- arn:aws:lambda:ap-northeast-2:845288260483:function:ACS-PassThrough
- arn:aws:lambda:eu-west-2:487402164563:function:ACS-PassThrough
- arn:aws:lambda:ap-southeast-1:377565633583:function:ACS-PassThrough
- arn:aws:lambda:ca-central-1:918755190332:function:ACS-PassThrough

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:lambda:[a-z0-9\-]*:[0-9]{12}:function:.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AppDetails

Service: Amazon SageMaker Service

Details about an Amazon SageMaker AI app.

Contents

AppName

The name of the app.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

AppType

The type of app.

Type: String

Valid Values: `JupyterServer | KernelGateway | DetailedProfiler | TensorBoard | CodeEditor | JupyterLab | RStudioServerPro | RSessionGateway | Canvas`

Required: No

CreationTime

The creation time.

Type: Timestamp

Required: No

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: No

ResourceSpec

Specifies the ARN's of a SageMaker AI image and SageMaker AI image version, and the instance type that the version runs on.

> **Note**
>
> When both `SageMakerImageVersionArn` and `SageMakerImageArn` are passed, `SageMakerImageVersionArn` is used. Any updates to `SageMakerImageArn` will not take effect if `SageMakerImageVersionArn` already exists in the `ResourceSpec` because `SageMakerImageVersionArn` always takes precedence. To clear the value set for `SageMakerImageVersionArn`, pass `None` as the value.

Type: ResourceSpec object

Required: No

SpaceName

The name of the space.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Status

The status.

Type: String

Valid Values: `Deleted | Deleting | Failed | InService | Pending`

Required: No

UserProfileName

The user profile name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AppImageConfigDetails

Service: Amazon SageMaker Service

The configuration for running a SageMaker AI image as a KernelGateway app.

Contents

AppImageConfigArn

The ARN of the AppImageConfig.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:app-image-config/.*`

Required: No

AppImageConfigName

The name of the AppImageConfig. Must be unique to your account.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

CodeEditorAppImageConfig

The configuration for the file system and the runtime, such as the environment variables and entry point.

Type: CodeEditorAppImageConfig object

Required: No

CreationTime

When the AppImageConfig was created.

Type: Timestamp

Required: No

JupyterLabAppImageConfig

The configuration for the file system and the runtime, such as the environment variables and entry point.

Type: JupyterLabAppImageConfig object

Required: No

KernelGatewayImageConfig

The configuration for the file system and kernels in the SageMaker AI image.

Type: KernelGatewayImageConfig object

Required: No

LastModifiedTime

When the AppImageConfig was last modified.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AppLifecycleManagement

Service: Amazon SageMaker Service

Settings that are used to configure and manage the lifecycle of Amazon SageMaker Studio applications.

Contents

IdleSettings

Settings related to idle shutdown of Studio applications.

Type: IdleSettings object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AppSpecification

Service: Amazon SageMaker Service

Configuration to run a processing job in a specified container image.

Contents

ImageUri

The container image to be run by the processing job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

ContainerArguments

The arguments for a container used to run a processing job.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

ContainerEntrypoint

The entrypoint for a container used to run a processing job.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ArtifactSource

Service: Amazon SageMaker Service

A structure describing the source of an artifact.

Contents

SourceUri

The URI of the source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*`

Required: Yes

SourceTypes

A list of source types.

Type: Array of ArtifactSourceType objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ArtifactSourceType

Service: Amazon SageMaker Service

The ID and ID type of an artifact source.

Contents

SourceIdType

The type of ID.

Type: String

Valid Values: `MD5Hash | S3ETag | S3Version | Custom`

Required: Yes

Value

The ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ArtifactSummary

Service: Amazon SageMaker Service

Lists a summary of the properties of an artifact. An artifact represents a URI addressable object or data. Some examples are a dataset and a model.

Contents

ArtifactArn

The Amazon Resource Name (ARN) of the artifact.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:artifact/.*`

Required: No

ArtifactName

The name of the artifact.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

ArtifactType

The type of the artifact.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

CreationTime

When the artifact was created.

Type: Timestamp

Required: No

LastModifiedTime

When the artifact was last modified.

Type: Timestamp

Required: No

Source

The source of the artifact.

Type: ArtifactSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AssociationInfo

Service: Amazon SageMaker Service

The data type used to describe the relationship between different sources.

Contents

DestinationArn

The Amazon Resource Name (ARN) of the `AssociationInfo` destination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

SourceArn

The Amazon Resource Name (ARN) of the `AssociationInfo` source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AssociationSummary

Service: Amazon SageMaker Service

Lists a summary of the properties of an association. An association is an entity that links other lineage or experiment entities. An example would be an association between a training job and a model.

Contents

AssociationType

The type of the association.

Type: String

Valid Values: `ContributedTo | AssociatedWith | DerivedFrom | Produced | SameAs`

Required: No

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

CreationTime

When the association was created.

Type: Timestamp

Required: No

DestinationArn

The Amazon Resource Name (ARN) of the destination.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: No

DestinationName

The name of the destination.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

DestinationType

The destination type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

SourceArn

The ARN of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: No

SourceName

The name of the source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

SourceType

The source type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AsyncInferenceClientConfig

Service: Amazon SageMaker Service

Configures the behavior of the client used by SageMaker to interact with the model container during asynchronous inference.

Contents

MaxConcurrentInvocationsPerInstance

The maximum number of concurrent requests sent by the SageMaker client to the model container. If no value is provided, SageMaker chooses an optimal value.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AsyncInferenceConfig

Service: Amazon SageMaker Service

Specifies configuration for how an endpoint performs asynchronous inference.

Contents

OutputConfig

Specifies the configuration for asynchronous inference invocation outputs.

Type: AsyncInferenceOutputConfig object

Required: Yes

ClientConfig

Configures the behavior of the client used by SageMaker to interact with the model container during asynchronous inference.

Type: AsyncInferenceClientConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AsyncInferenceNotificationConfig

Service: Amazon SageMaker Service

Specifies the configuration for notifications of inference results for asynchronous inference.

Contents

ErrorTopic

Amazon SNS topic to post a notification to when inference fails. If no topic is provided, no notification is sent on failure.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sns:[a-z0-9\-]*:[0-9]{12}:[a-zA-Z0-9_.-]+`

Required: No

IncludeInferenceResponseIn

The Amazon SNS topics where you want the inference response to be included.

> **Note**
>
> The inference response is included only if the response size is less than or equal to 128 KB.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2 items.

Valid Values: `SUCCESS_NOTIFICATION_TOPIC | ERROR_NOTIFICATION_TOPIC`

Required: No

SuccessTopic

Amazon SNS topic to post a notification to when inference completes successfully. If no topic is provided, no notification is sent on success.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sns:[a-z0-9\-]*:[0-9]{12}:[a-zA-Z0-9_.-]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AsyncInferenceOutputConfig

Service: Amazon SageMaker Service

Specifies the configuration for asynchronous inference invocation outputs.

Contents

KmsKeyId

The AWS Key Management Service (AWS KMS) key that SageMaker uses to encrypt the asynchronous inference output in Amazon S3.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

NotificationConfig

Specifies the configuration for notifications of inference results for asynchronous inference.

Type: AsyncInferenceNotificationConfig object

Required: No

S3FailurePath

The Amazon S3 location to upload failure inference responses to.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `(https|s3)://([^/])/?(.*)`

Required: No

S3OutputPath

The Amazon S3 location to upload inference responses to.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `(https|s3)://([^/])/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AthenaDatasetDefinition

Service: Amazon SageMaker Service

Configuration for Athena Dataset Definition input.

Contents

Catalog

The name of the data catalog used in Athena query execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\t]*`

Required: Yes

Database

The name of the database used in the Athena query execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.*`

Required: Yes

OutputFormat

The data storage format for Athena query results.

Type: String

Valid Values: `PARQUET | ORC | AVRO | JSON | TEXTFILE`

Required: Yes

OutputS3Uri

The location in Amazon S3 where Athena query results are stored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

QueryString

The SQL query statements, to be executed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 4096.

Pattern: `[\s\S]+`

Required: Yes

KmsKeyId

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt data generated from an Athena query execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

OutputCompression

The compression used for Athena query results.

Type: String

Valid Values: `GZIP | SNAPPY | ZLIB`

Required: No

WorkGroup

The name of the workgroup in which the Athena query is being started.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9._-]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AuthorizedUrl

Service: Amazon SageMaker Service

Contains a presigned URL and its associated local file path for downloading hub content artifacts.

Contents

LocalPath

The recommended local file path where the downloaded file should be stored to maintain proper directory structure and file organization.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `\/.*`

Required: No

Url

The presigned S3 URL that provides temporary, secure access to download the file. URLs expire within 15 minutes for security purposes.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLAlgorithmConfig

Service: Amazon SageMaker Service

The selection of algorithms trained on your dataset to generate the model candidates for an Autopilot job.

Contents

AutoMLAlgorithms

The selection of algorithms trained on your dataset to generate the model candidates for an Autopilot job.

- **For the tabular problem type `TabularJobConfig`:**

  > **Note**
  >
  > Selected algorithms must belong to the list corresponding to the training mode set in AutoMLJobConfig.Mode (`ENSEMBLING` or `HYPERPARAMETER_TUNING`). Choose a minimum of 1 algorithm.

  - In `ENSEMBLING` mode:
    - "catboost"
    - "extra-trees"
    - "fastai"
    - "lightgbm"
    - "linear-learner"
    - "nn-torch"
    - "randomforest"
    - "xgboost"
  - In `HYPERPARAMETER_TUNING` mode:
    - "linear-learner"
    - "mlp"
    - "xgboost"
- **For the time-series forecasting problem type `TimeSeriesForecastingJobConfig`:**

- Choose your algorithms from this list.
  - "cnn-qr"
  - "deepar"
  - "prophet"
  - "arima"
  - "npts"
  - "ets"

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 11 items.

Valid Values: `xgboost | linear-learner | mlp | lightgbm | catboost | randomforest | extra-trees | nn-torch | fastai | cnn-qr | deepar | prophet | npts | arima | ets`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLCandidate

Service: Amazon SageMaker Service

Information about a candidate produced by an AutoML training job, including its status, steps, and other properties.

Contents

CandidateName

The name of the candidate.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Required: Yes

CandidateStatus

The candidate's status.

Type: String

Valid Values: `Completed | InProgress | Failed | Stopped | Stopping`

Required: Yes

CandidateSteps

Information about the candidate's steps.

Type: Array of AutoMLCandidateStep objects

Required: Yes

CreationTime

The creation time.

Type: Timestamp

Required: Yes

LastModifiedTime

The last modified time.

Type: Timestamp

Required: Yes

ObjectiveStatus

The objective's status.

Type: String

Valid Values: `Succeeded | Pending | Failed`

Required: Yes

CandidateProperties

The properties of an AutoML candidate job.

Type: CandidateProperties object

Required: No

EndTime

The end time.

Type: Timestamp

Required: No

FailureReason

The failure reason.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

FinalAutoMLJobObjectiveMetric

The best candidate result from an AutoML training job.

Type: FinalAutoMLJobObjectiveMetric object

Required: No

InferenceContainerDefinitions

The mapping of all supported processing unit (CPU, GPU, etc...) to inference container definitions for the candidate. This field is populated for the AutoML jobs V2 (for example, for jobs created by calling `CreateAutoMLJobV2`) related to image or text classification problem types only.

Type: String to array of AutoMLContainerDefinition objects map

Map Entries: Minimum number of 0 items. Maximum number of 2 items.

Valid Keys: `CPU | GPU`

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

InferenceContainers

Information about the recommended inference container definitions.

Type: Array of AutoMLContainerDefinition objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLCandidateGenerationConfig

Service: Amazon SageMaker Service

Stores the configuration information for how a candidate is generated (optional).

Contents

AlgorithmsConfig

Stores the configuration information for the selection of algorithms trained on tabular data.

The list of available algorithms to choose from depends on the training mode set in `TabularJobConfig.Mode`.

- `AlgorithmsConfig` should not be set if the training mode is set on `AUTO`.
- When `AlgorithmsConfig` is provided, one `AutoMLAlgorithms` attribute must be set and one only.

  If the list of algorithms provided as values for `AutoMLAlgorithms` is empty, `CandidateGenerationConfig` uses the full set of algorithms for the given training mode.
- When `AlgorithmsConfig` is not provided, `CandidateGenerationConfig` uses the full set of algorithms for the given training mode.

For the list of all algorithms per problem type and training mode, see AutoMLAlgorithmConfig.

For more information on each algorithm, see the Algorithm support section in Autopilot developer guide.

Type: Array of AutoMLAlgorithmConfig objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

FeatureSpecificationS3Uri

A URL to the Amazon S3 data source containing selected features from the input data source to run an Autopilot job. You can input `FeatureAttributeNames` (optional) in JSON format as shown below:

```
{ "FeatureAttributeNames":["col1", "col2", ...] }.
```

You can also specify the data type of the feature (optional) in the format shown below:

`{ "FeatureDataTypes":{"col1":"numeric", "col2":"categorical" ... } }`

> **Note**
>
> These column keys may not include the target column.

In ensembling mode, Autopilot only supports the following data types: `numeric`, `categorical`, `text`, and `datetime`. In HPO mode, Autopilot can support `numeric`, `categorical`, `text`, `datetime`, and `sequence`.

If only `FeatureDataTypes` is provided, the column keys (`col1`, `col2`,..) should be a subset of the column names in the input data.

If both `FeatureDataTypes` and `FeatureAttributeNames` are provided, then the column keys should be a subset of the column names provided in `FeatureAttributeNames`.

The key name `FeatureAttributeNames` is fixed. The values listed in `["col1", "col2", ...]` are case sensitive and should be a list of strings containing unique values that are a subset of the column names in the input data. The list of columns provided must not include the target column.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLCandidateStep

Service: Amazon SageMaker Service

Information about the steps for a candidate and what step it is working on.

Contents

CandidateStepArn

The ARN for the candidate's step.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:.*/.*`

Required: Yes

CandidateStepName

The name for the candidate's step.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Required: Yes

CandidateStepType

Whether the candidate is at the transform, training, or processing step.

Type: String

Valid Values: `AWS::SageMaker::TrainingJob | AWS::SageMaker::TransformJob | AWS::SageMaker::ProcessingJob`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLChannel

Service: Amazon SageMaker Service

A channel is a named input source that training algorithms can consume. The validation dataset size is limited to less than 2 GB. The training dataset size must be less than 100 GB. For more information, see Channel.

Note

A validation dataset must contain the same headers as the training dataset.

Contents

TargetAttributeName

The name of the target variable in supervised learning, usually represented by 'y'.

Type: String

Length Constraints: Minimum length of 1.

Required: Yes

ChannelType

The channel type (optional) is an enum string. The default value is `training`. Channels for training and validation must share the same `ContentType` and `TargetAttributeName`. For information on specifying training and validation channel types, see How to specify training and validation datasets.

Type: String

Valid Values: `training | validation`

Required: No

CompressionType

You can use `Gzip` or None. The default value is None.

Type: String

Valid Values: `None | Gzip`

Required: No

ContentType

The content type of the data from the input source. You can use `text/csv;header=present` or `x-application/vnd.amazon+parquet`. The default value is `text/csv;header=present`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

DataSource

The data source for an AutoML channel.

Type: AutoMLDataSource object

Required: No

SampleWeightAttributeName

If specified, this column name indicates which column of the dataset should be treated as sample weights for use by the objective metric during the training, evaluation, and the selection of the best model. This column is not considered as a predictive feature. For more information on Autopilot metrics, see Metrics and validation.

Sample weights should be numeric, non-negative, with larger values indicating which rows are more important than others. Data points that have invalid or no weight value are excluded.

Support for sample weights is available in Ensembling mode only.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9_-]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLComputeConfig

Service: Amazon SageMaker Service

> **Note**
>
> This data type is intended for use exclusively by SageMaker Canvas and cannot be used in other contexts at the moment.

Specifies the compute configuration for an AutoML job V2.

Contents

EmrServerlessComputeConfig

The configuration for using EMR Serverless to run the AutoML job V2.

To allow your AutoML job V2 to automatically initiate a remote job on EMR Serverless when additional compute resources are needed to process large datasets, you need to provide an `EmrServerlessComputeConfig` object, which includes an `ExecutionRoleARN` attribute, to the `AutoMLComputeConfig` of the AutoML job V2 input request.

By seamlessly transitioning to EMR Serverless when required, the AutoML job can handle datasets that would otherwise exceed the initially provisioned resources, without any manual intervention from you.

EMR Serverless is available for the tabular and time series problem types. We recommend setting up this option for tabular datasets larger than 5 GB and time series datasets larger than 30 GB.

Type: EmrServerlessComputeConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLContainerDefinition

Service: Amazon SageMaker Service

A list of container definitions that describe the different containers that make up an AutoML candidate. For more information, see ContainerDefinition.

Contents

Image

The Amazon Elastic Container Registry (Amazon ECR) path of the container. For more information, see ContainerDefinition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[\S]+`

Required: Yes

ModelDataUrl

The location of the model artifacts. For more information, see ContainerDefinition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

Environment

The environment variables to set in the container. For more information, see ContainerDefinition.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 1024.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 1024.

Value Pattern: `[\S\s]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLDataSource

Service: Amazon SageMaker Service

The data source for the Autopilot job.

Contents

S3DataSource

The Amazon S3 location of the input data.

Type: AutoMLS3DataSource object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLDataSplitConfig

Service: Amazon SageMaker Service

This structure specifies how to split the data into train and validation datasets.

The validation and training datasets must contain the same headers. For jobs created by calling `CreateAutoMLJob`, the validation dataset must be less than 2 GB in size.

Contents

ValidationFraction

The validation fraction (optional) is a float that specifies the portion of the training dataset to be used for validation. The default value is 0.2, and values must be greater than 0 and less than 1. We recommend setting this value to be less than 0.5.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLJobArtifacts

Service: Amazon SageMaker Service

The artifacts that are generated during an AutoML job.

Contents

CandidateDefinitionNotebookLocation

The URL of the notebook location.

Type: String

Length Constraints: Minimum length of 1.

Required: No

DataExplorationNotebookLocation

The URL of the notebook location.

Type: String

Length Constraints: Minimum length of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLJobChannel

Service: Amazon SageMaker Service

A channel is a named input source that training algorithms can consume. This channel is used for AutoML jobs V2 (jobs created by calling CreateAutoMLJobV2).

Contents

ChannelType

The type of channel. Defines whether the data are used for training or validation. The default value is `training`. Channels for `training` and `validation` must share the same `ContentType`

> **Note**
>
> The type of channel defaults to `training` for the time-series forecasting problem type.

Type: String

Valid Values: `training | validation`

Required: No

CompressionType

The allowed compression types depend on the input format and problem type. We allow the compression type `Gzip` for `S3Prefix` inputs on tabular data only. For all other inputs, the compression type should be `None`. If no compression type is provided, we default to `None`.

Type: String

Valid Values: `None | Gzip`

Required: No

ContentType

The content type of the data from the input source. The following are the allowed content types for different problems:

- For tabular problem types: `text/csv;header=present` or `x-application/vnd.amazon+parquet`. The default value is `text/csv;header=present`.
- For image classification: `image/png`, `image/jpeg`, or `image/*`. The default value is `image/*`.
- For text classification: `text/csv;header=present` or `x-application/vnd.amazon+parquet`. The default value is `text/csv;header=present`.
- For time-series forecasting: `text/csv;header=present` or `x-application/vnd.amazon+parquet`. The default value is `text/csv;header=present`.
- For text generation (LLMs fine-tuning): `text/csv;header=present` or `x-application/vnd.amazon+parquet`. The default value is `text/csv;header=present`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

DataSource

The data source for an AutoML channel (Required).

Type: AutoMLDataSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLJobCompletionCriteria

Service: Amazon SageMaker Service

How long a job is allowed to run, or how many candidates a job is allowed to generate.

Contents

MaxAutoMLJobRuntimeInSeconds

The maximum runtime, in seconds, an AutoML job has to complete.

If an AutoML job exceeds the maximum runtime, the job is stopped automatically and its processing is ended gracefully. The AutoML job identifies the best model whose training was completed and marks it as the best-performing model. Any unfinished steps of the job, such as automatic one-click Autopilot model deployment, are not completed.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

MaxCandidates

The maximum number of times a training job is allowed to run.

For text and image classification, time-series forecasting, as well as text generation (LLMs fine-tuning) problem types, the supported value is 1. For tabular problem types, the maximum value is 750.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 750.

Required: No

MaxRuntimePerTrainingJobInSeconds

The maximum time, in seconds, that each training job executed inside hyperparameter tuning is allowed to run as part of a hyperparameter tuning job. For more information, see the StoppingCondition used by the CreateHyperParameterTuningJob action.

For job V2s (jobs created by calling `CreateAutoMLJobV2`), this field controls the runtime of the job candidate.

For TextGenerationJobConfig problem types, the maximum time defaults to 72 hours (259200 seconds).

Type: Integer

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLJobConfig

Service: Amazon SageMaker Service

A collection of settings used for an AutoML job.

Contents

CandidateGenerationConfig

The configuration for generating a candidate for an AutoML job (optional).

Type: AutoMLCandidateGenerationConfig object

Required: No

CompletionCriteria

How long an AutoML job is allowed to run, or how many candidates a job is allowed to generate.

Type: AutoMLJobCompletionCriteria object

Required: No

DataSplitConfig

The configuration for splitting the input training dataset.

Type: AutoMLDataSplitConfig

Type: AutoMLDataSplitConfig object

Required: No

Mode

The method that Autopilot uses to train the data. You can either specify the mode manually or let Autopilot choose for you based on the dataset size by selecting `AUTO`. In `AUTO` mode, Autopilot chooses `ENSEMBLING` for datasets smaller than 100 MB, and `HYPERPARAMETER_TUNING` for larger ones.

The `ENSEMBLING` mode uses a multi-stack ensemble model to predict classification and regression tasks directly from your dataset. This machine learning mode combines several base models to produce an optimal predictive model. It then uses a stacking ensemble method to

combine predictions from contributing members. A multi-stack ensemble model can provide better performance over a single model by combining the predictive capabilities of multiple models. See Autopilot algorithm support for a list of algorithms supported by `ENSEMBLING` mode.

The `HYPERPARAMETER_TUNING` (HPO) mode uses the best hyperparameters to train the best version of a model. HPO automatically selects an algorithm for the type of problem you want to solve. Then HPO finds the best hyperparameters according to your objective metric. See Autopilot algorithm support for a list of algorithms supported by `HYPERPARAMETER_TUNING` mode.

Type: String

Valid Values: `AUTO | ENSEMBLING | HYPERPARAMETER_TUNING`

Required: No

SecurityConfig

The security configuration for traffic encryption or Amazon VPC settings.

Type: AutoMLSecurityConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLJobObjective

Service: Amazon SageMaker Service

Specifies a metric to minimize or maximize as the objective of an AutoML job.

Contents

MetricName

The name of the objective metric used to measure the predictive quality of a machine learning system. During training, the model's parameters are updated iteratively to optimize its performance based on the feedback provided by the objective metric when evaluating the model on the validation dataset.

The list of available metrics supported by Autopilot and the default metric applied when you do not specify a metric name explicitly depend on the problem type.

- For tabular problem types:
  - List of available metrics:
    - Regression: `MAE`, `MSE`, `R2`, `RMSE`
    - Binary classification: `Accuracy`, `AUC`, `BalancedAccuracy`, `F1`, `Precision`, `Recall`
    - Multiclass classification: `Accuracy`, `BalancedAccuracy`, `F1macro`, `PrecisionMacro`, `RecallMacro`

    For a description of each metric, see Autopilot metrics for classification and regression.
  - Default objective metrics:
    - Regression: `MSE`.
    - Binary classification: `F1`.
    - Multiclass classification: `Accuracy`.
- For image or text classification problem types:
  - List of available metrics: `Accuracy`

    For a description of each metric, see Autopilot metrics for text and image classification.
  - Default objective metrics: `Accuracy`
- For time-series forecasting problem types:
  - List of available metrics: `RMSE`, `wQL`, `Average wQL`, `MASE`, `MAPE`, `WAPE`

For a description of each metric, see Autopilot metrics for time-series forecasting.

- Default objective metrics: `AverageWeightedQuantileLoss`

- For text generation problem types (LLMs fine-tuning): Fine-tuning language models in Autopilot does not require setting the `AutoMLJobObjective` field. Autopilot fine-tunes LLMs without requiring multiple candidates to be trained and evaluated. Instead, using your dataset, Autopilot directly fine-tunes your target model to enhance a default objective metric, the cross-entropy loss. After fine-tuning a language model, you can evaluate the quality of its generated text using different metrics. For a list of the available metrics, see Metrics for fine-tuning LLMs in Autopilot.

Type: String

Valid Values: `Accuracy | MSE | F1 | F1macro | AUC | RMSE | BalancedAccuracy | R2 | Recall | RecallMacro | Precision | PrecisionMacro | MAE | MAPE | MASE | WAPE | AverageWeightedQuantileLoss`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLJobStepMetadata

Service: Amazon SageMaker Service

Metadata for an AutoML job step.

Contents

Arn

The Amazon Resource Name (ARN) of the AutoML job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLJobSummary

Service: Amazon SageMaker Service

Provides a summary about an AutoML job.

Contents

AutoMLJobArn

The ARN of the AutoML job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

Required: Yes

AutoMLJobName

The name of the AutoML job you are requesting.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

AutoMLJobSecondaryStatus

The secondary status of the AutoML job.

Type: String

Valid Values: `Starting | MaxCandidatesReached | Failed | Stopped | MaxAutoMLJobRuntimeReached | Stopping | CandidateDefinitionsGenerated | Completed | ExplainabilityError | DeployingModel | ModelDeploymentError | GeneratingModelInsightsReport | ModelInsightsError | AnalyzingData | FeatureEngineering | ModelTuning | GeneratingExplainabilityReport | TrainingModels | PreTraining`

Required: Yes

AutoMLJobStatus

The status of the AutoML job.

Type: String

Valid Values: `Completed | InProgress | Failed | Stopped | Stopping`

Required: Yes

CreationTime

When the AutoML job was created.

Type: Timestamp

Required: Yes

LastModifiedTime

When the AutoML job was last modified.

Type: Timestamp

Required: Yes

EndTime

The end time of an AutoML job.

Type: Timestamp

Required: No

FailureReason

The failure reason of an AutoML job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

PartialFailureReasons

The list of reasons for partial failures within an AutoML job.

Type: Array of AutoMLPartialFailureReason objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLOutputDataConfig

Service: Amazon SageMaker Service

The output data configuration.

Contents

S3OutputPath

The Amazon S3 output path. Must be 512 characters or less.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

KmsKeyId

The AWS Key Management Service encryption key ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLPartialFailureReason

Service: Amazon SageMaker Service

The reason for a partial failure of an AutoML job.

Contents

PartialFailureMessage

The message containing the reason for a partial failure of an AutoML job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLProblemTypeConfig

Service: Amazon SageMaker Service

A collection of settings specific to the problem type used to configure an AutoML job V2. There must be one and only one config of the following type.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

ImageClassificationJobConfig

Settings used to configure an AutoML job V2 for the image classification problem type.

Type: ImageClassificationJobConfig object

Required: No

TabularJobConfig

Settings used to configure an AutoML job V2 for the tabular problem type (regression, classification).

Type: TabularJobConfig object

Required: No

TextClassificationJobConfig

Settings used to configure an AutoML job V2 for the text classification problem type.

Type: TextClassificationJobConfig object

Required: No

TextGenerationJobConfig

Settings used to configure an AutoML job V2 for the text generation (LLMs fine-tuning) problem type.

> **Note**
>
> The text generation models that support fine-tuning in Autopilot are currently accessible exclusively in regions supported by Canvas. Refer to the documentation of Canvas for the full list of its supported Regions.

Type: TextGenerationJobConfig object

Required: No

TimeSeriesForecastingJobConfig

Settings used to configure an AutoML job V2 for the time-series forecasting problem type.

Type: TimeSeriesForecastingJobConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLProblemTypeResolvedAttributes

Service: Amazon SageMaker Service

Stores resolved attributes specific to the problem type of an AutoML job V2.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

TabularResolvedAttributes

The resolved attributes for the tabular problem type.

Type: TabularResolvedAttributes object

Required: No

TextGenerationResolvedAttributes

The resolved attributes for the text generation problem type.

Type: TextGenerationResolvedAttributes object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLResolvedAttributes

Service: Amazon SageMaker Service

The resolved attributes used to configure an AutoML job V2.

Contents

AutoMLJobObjective

Specifies a metric to minimize or maximize as the objective of an AutoML job.

Type: AutoMLJobObjective object

Required: No

AutoMLProblemTypeResolvedAttributes

Defines the resolved attributes specific to a problem type.

Type: AutoMLProblemTypeResolvedAttributes object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

CompletionCriteria

How long a job is allowed to run, or how many candidates a job is allowed to generate.

Type: AutoMLJobCompletionCriteria object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLS3DataSource

Service: Amazon SageMaker Service

Describes the Amazon S3 data source.

Contents

S3DataType

The data type.

- If you choose `S3Prefix`, `S3Uri` identifies a key name prefix. SageMaker AI uses all objects that match the specified key name prefix for model training.

  The `S3Prefix` should have the following format:

  `s3://DOC-EXAMPLE-BUCKET/DOC-EXAMPLE-FOLDER-OR-FILE`

- If you choose `ManifestFile`, `S3Uri` identifies an object that is a manifest file containing a list of object keys that you want SageMaker AI to use for model training.

  A `ManifestFile` should have the format shown below:

  `[ {"prefix": "s3://DOC-EXAMPLE-BUCKET/DOC-EXAMPLE-FOLDER/DOC-EXAMPLE-PREFIX/"},`

  `"DOC-EXAMPLE-RELATIVE-PATH/DOC-EXAMPLE-FOLDER/DATA-1",`

  `"DOC-EXAMPLE-RELATIVE-PATH/DOC-EXAMPLE-FOLDER/DATA-2",`

  `... "DOC-EXAMPLE-RELATIVE-PATH/DOC-EXAMPLE-FOLDER/DATA-N" ]`

- If you choose `AugmentedManifestFile`, `S3Uri` identifies an object that is an augmented manifest file in JSON lines format. This file contains the data you want to use for model training. `AugmentedManifestFile` is available for V2 API jobs only (for example, for jobs created by calling `CreateAutoMLJobV2`).

  Here is a minimal, single-record example of an `AugmentedManifestFile`:

  `{"source-ref": "s3://DOC-EXAMPLE-BUCKET/DOC-EXAMPLE-FOLDER/cats/cat.jpg",`

  `"label-metadata": {"class-name": "cat"}`

For more information on `AugmentedManifestFile`, see Provide Dataset Metadata to Training Jobs with an Augmented Manifest File.

Type: String

Valid Values: `ManifestFile | S3Prefix | AugmentedManifestFile`

Required: Yes

S3Uri

The URL to the Amazon S3 data source. The Uri refers to the Amazon S3 prefix or ManifestFile depending on the data type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoMLSecurityConfig

Service: Amazon SageMaker Service

Security options.

Contents

EnableInterContainerTrafficEncryption

Whether to use traffic encryption between the container layers.

Type: Boolean

Required: No

VolumeKmsKeyId

The key used to encrypt stored data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

VpcConfig

The VPC configuration.

Type: VpcConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoParameter

Service: Amazon SageMaker Service

The name and an example value of the hyperparameter that you want to use in Autotune. If Automatic model tuning (AMT) determines that your hyperparameter is eligible for Autotune, an optimal hyperparameter range is selected for you.

Contents

Name

The name of the hyperparameter to optimize using Autotune.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

ValueHint

An example value of the hyperparameter to optimize using Autotune.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoRollbackConfig

Service: Amazon SageMaker Service

Automatic rollback configuration for handling endpoint deployment failures and recovery.

Contents

Alarms

List of CloudWatch alarms in your account that are configured to monitor metrics on an endpoint. If any alarms are tripped during a deployment, SageMaker rolls back the deployment.

Type: Array of Alarm objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Autotune

Service: Amazon SageMaker Service

A flag to indicate if you want to use Autotune to automatically find optimal values for the following fields:

- ParameterRanges: The names and ranges of parameters that a hyperparameter tuning job can optimize.
- ResourceLimits: The maximum resources that can be used for a training job. These resources include the maximum number of training jobs, the maximum runtime of a tuning job, and the maximum number of training jobs to run at the same time.
- TrainingJobEarlyStoppingType: A flag that specifies whether or not to use early stopping for training jobs launched by a hyperparameter tuning job.
- RetryStrategy: The number of times to retry a training job.
- Strategy: Specifies how hyperparameter tuning chooses the combinations of hyperparameter values to use for the training jobs that it launches.
- ConvergenceDetected: A flag to indicate that Automatic model tuning (AMT) has detected model convergence.

Contents

Mode

Set `Mode` to `Enabled` if you want to use Autotune.

Type: String

Valid Values: `Enabled`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

AvailableUpgrade

Service: Amazon SageMaker Service

Contains information about an available upgrade for a SageMaker Partner AI App, including the version number and release notes.

Contents

ReleaseNotes

A list of release notes describing the changes and improvements included in the available upgrade version.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

Version

The semantic version number of the available upgrade for the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `\d+\.\d+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BaseModel

Service: Amazon SageMaker Service

Identifies the foundation model that was used as the starting point for model customization.

Contents

HubContentName

The hub content name of the base model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

HubContentVersion

The hub content version of the base model.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: No

RecipeName

The recipe name of the base model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchAddClusterNodesError

Service: Amazon SageMaker Service

Information about an error that occurred during the node addition operation.

Contents

ErrorCode

The error code associated with the failure. Possible values include `InstanceGroupNotFound` and `InvalidInstanceGroupState`.

Type: String

Valid Values: `InstanceGroupNotFound | InvalidInstanceGroupStatus | IncompatibleAvailabilityZones | IncompatibleInstanceTypes`

Required: Yes

FailedCount

The number of nodes that failed to be added to the specified instance group.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

InstanceGroupName

The name of the instance group for which the error occurred.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.+`

Required: Yes

AvailabilityZones

The availability zones associated with the failed node addition request.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Pattern: `[a-z]{2}-[a-z]+-\d[a-z]`

Required: No

InstanceTypes

The instance types associated with the failed node addition request.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge |`

`ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge`

Required: No

Message

A descriptive message providing additional details about the error.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchDataCaptureConfig

Service: Amazon SageMaker Service

Configuration to control how SageMaker captures inference data for batch transform jobs.

Contents

DestinationS3Uri

The Amazon S3 location being used to capture the data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

GenerateInferenceId

Flag that indicates whether to append inference id to the output.

Type: Boolean

Required: No

KmsKeyId

The Amazon Resource Name (ARN) of a AWS Key Management Service key that SageMaker uses to encrypt data on the storage volume attached to the ML compute instance that hosts the batch transform job.

The KmsKeyId can be any of the following formats:

- Key ID: `1234abcd-12ab-34cd-56ef-1234567890ab`
- Key ARN: `arn:aws:kms:us-west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab`
- Alias name: `alias/ExampleAlias`
- Alias name ARN: `arn:aws:kms:us-west-2:111122223333:alias/ExampleAlias`

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchDeleteClusterNodeLogicalIdsError

Service: Amazon SageMaker Service

Information about an error that occurred when attempting to delete a node identified by its `NodeLogicalId`.

Contents

Code

The error code associated with the failure. Possible values include `NodeLogicalIdNotFound`, `InvalidNodeStatus`, and `InternalError`.

Type: String

Valid Values: `NodeIdNotFound | InvalidNodeStatus | NodeIdInUse`

Required: Yes

Message

A descriptive message providing additional details about the error.

Type: String

Required: Yes

NodeLogicalId

The `NodeLogicalId` of the node that could not be deleted.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchDeleteClusterNodesError

Service: Amazon SageMaker Service

Represents an error encountered when deleting a node from a SageMaker HyperPod cluster.

Contents

Code

The error code associated with the error encountered when deleting a node.

The code provides information about the specific issue encountered, such as the node not being found, the node's status being invalid for deletion, or the node ID being in use by another process.

Type: String

Valid Values: `NodeIdNotFound | InvalidNodeStatus | NodeIdInUse`

Required: Yes

Message

A message describing the error encountered when deleting a node.

Type: String

Required: Yes

NodeId

The ID of the node that encountered an error during the deletion process.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchDescribeModelPackageError

Service: Amazon SageMaker Service

The error code and error description associated with the resource.

Contents

ErrorCode

Type: String

Required: Yes

ErrorResponse

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchDescribeModelPackageSummary

Service: Amazon SageMaker Service

Provides summary information about the model package.

Contents

CreationTime

The creation time of the mortgage package summary.

Type: Timestamp

Required: Yes

InferenceSpecification

Defines how to perform inference generation after a training job is run.

Type: InferenceSpecification object

Required: Yes

ModelPackageArn

The Amazon Resource Name (ARN) of the model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: Yes

ModelPackageGroupName

The group name for the model package

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelPackageStatus

The status of the mortgage package.

Type: String

Valid Values: `Pending | InProgress | Completed | Failed | Deleting`

Required: Yes

ModelApprovalStatus

The approval status of the model.

Type: String

Valid Values: `Approved | Rejected | PendingManualApproval`

Required: No

ModelPackageDescription

The description of the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

ModelPackageRegistrationType

The package registration type of the model package summary.

Type: String

Valid Values: `Logged | Registered`

Required: No

ModelPackageVersion

The version number of a versioned model.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchRebootClusterNodeLogicalIdsError

Service: Amazon SageMaker Service

Represents an error encountered when rebooting a node (identified by its logical node ID) from a SageMaker HyperPod cluster.

Contents

ErrorCode

The error code associated with the error encountered when rebooting a node by logical node ID.

Possible values:

- `InstanceIdNotFound`: The node does not exist in the specified cluster.
- `InvalidInstanceStatus`: The node is in a state that does not allow rebooting. Wait for the node to finish any ongoing changes before retrying.
- `InstanceIdInUse`: Another operation is already in progress for this node. Wait for the operation to complete before retrying.
- `InternalServerError`: An internal error occurred while processing this node.

Type: String

Valid Values: `InstanceIdNotFound | InvalidInstanceStatus | InstanceIdInUse | InternalServerError`

Required: Yes

Message

A human-readable message describing the error encountered when rebooting a node by logical node ID.

Type: String

Required: Yes

NodeLogicalId

The logical node ID of the node that encountered an error during the reboot operation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchRebootClusterNodesError

Service: Amazon SageMaker Service

Represents an error encountered when rebooting a node from a SageMaker HyperPod cluster.

Contents

ErrorCode

The error code associated with the error encountered when rebooting a node.

Possible values:

- `InstanceIdNotFound`: The instance does not exist in the specified cluster.
- `InvalidInstanceStatus`: The instance is in a state that does not allow rebooting. Wait for the instance to finish any ongoing changes before retrying.
- `InstanceIdInUse`: Another operation is already in progress for this node. Wait for the operation to complete before retrying.
- `InternalServerError`: An internal error occurred while processing this node.

Type: String

Valid Values: `InstanceIdNotFound | InvalidInstanceStatus | InstanceIdInUse | InternalServerError`

Required: Yes

Message

A human-readable message describing the error encountered when rebooting a node.

Type: String

Required: Yes

NodeId

The EC2 instance ID of the node that encountered an error during the reboot operation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchReplaceClusterNodeLogicalIdsError

Service: Amazon SageMaker Service

Represents an error encountered when replacing a node (identified by its logical node ID) in a SageMaker HyperPod cluster.

Contents

ErrorCode

The error code associated with the error encountered when replacing a node by logical node ID.

Possible values:

- `InstanceIdNotFound`: The node does not exist in the specified cluster.
- `InvalidInstanceStatus`: The node is in a state that does not allow replacement. Wait for the node to finish any ongoing changes before retrying.
- `InstanceIdInUse`: Another operation is already in progress for this node. Wait for the operation to complete before retrying.
- `InternalServerError`: An internal error occurred while processing this node.

Type: String

Valid Values: `InstanceIdNotFound | InvalidInstanceStatus | InstanceIdInUse | InternalServerError`

Required: Yes

Message

A human-readable message describing the error encountered when replacing a node by logical node ID.

Type: String

Required: Yes

NodeLogicalId

The logical node ID of the node that encountered an error during the replacement operation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchReplaceClusterNodesError

Service: Amazon SageMaker Service

Represents an error encountered when replacing a node in a SageMaker HyperPod cluster.

Contents

ErrorCode

The error code associated with the error encountered when replacing a node.

Possible values:

- `InstanceIdNotFound`: The instance does not exist in the specified cluster.
- `InvalidInstanceStatus`: The instance is in a state that does not allow replacement. Wait for the instance to finish any ongoing changes before retrying.
- `InstanceIdInUse`: Another operation is already in progress for this node. Wait for the operation to complete before retrying.
- `InternalServerError`: An internal error occurred while processing this node.

Type: String

Valid Values: `InstanceIdNotFound | InvalidInstanceStatus | InstanceIdInUse | InternalServerError`

Required: Yes

Message

A human-readable message describing the error encountered when replacing a node.

Type: String

Required: Yes

NodeId

The EC2 instance ID of the node that encountered an error during the replacement operation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchTransformInput

Service: Amazon SageMaker Service

Input object for the batch transform job.

Contents

DataCapturedDestinationS3Uri

The Amazon S3 location being used to capture the data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `(https|s3)://([^/])/?(.*)`

Required: Yes

DatasetFormat

The dataset format for your batch transform job.

Type: MonitoringDatasetFormat object

Required: Yes

LocalPath

Path to the filesystem where the batch transform data is available to the container.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

EndTimeOffset

If specified, monitoring jobs subtract this time from the end time. For information about using offsets for scheduling monitoring jobs, see Schedule Model Quality Monitoring Jobs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 15.

Pattern: `.?P.*`

Required: No

ExcludeFeaturesAttribute

The attributes of the input data to exclude from the analysis.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 100.

Required: No

FeaturesAttribute

The attributes of the input data that are the input features.

Type: String

Required: No

InferenceAttribute

The attribute of the input data that represents the ground truth label.

Type: String

Required: No

ProbabilityAttribute

In a classification problem, the attribute that represents the class probability.

Type: String

Required: No

ProbabilityThresholdAttribute

The threshold for the class probability to be evaluated as a positive result.

Type: Double

Required: No

S3DataDistributionType

Whether input data distributed in Amazon S3 is fully replicated or sharded by an S3 key. Defaults to `FullyReplicated`

Type: String

Valid Values: `FullyReplicated | ShardedByS3Key`

Required: No

S3InputMode

Whether the `Pipe` or `File` is used as the input mode for transferring data for the monitoring job. `Pipe` mode is recommended for large datasets. `File` mode is useful for small files that fit in memory. Defaults to `File`.

Type: String

Valid Values: `Pipe | File`

Required: No

StartTimeOffset

If specified, monitoring jobs substract this time from the start time. For information about using offsets for scheduling monitoring jobs, see Schedule Model Quality Monitoring Jobs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 15.

Pattern: `.?P.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

BedrockCustomModelDeploymentMetadata

Service: Amazon SageMaker Service

The metadata of the Amazon Bedrock custom model deployment.

Contents

Arn

The Amazon Resource Name (ARN) for the Amazon Bedrock custom model deployment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockCustomModelMetadata

Service: Amazon SageMaker Service

The metadata of the Amazon Bedrock custom model.

Contents

Arn

The Amazon Resource Name (ARN) of the Amazon Bedrock custom model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockModelImportMetadata

Service: Amazon SageMaker Service

The metadata of the Amazon Bedrock model import.

Contents

Arn

The Amazon Resource Name (ARN) of the Amazon Bedrock model import.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockProvisionedModelThroughputMetadata

Service: Amazon SageMaker Service

The metadata of the Amazon Bedrock provisioned model throughput.

Contents

Arn

The Amazon Resource Name (ARN) of the Amazon Bedrock provisioned model throughput.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BestObjectiveNotImproving

Service: Amazon SageMaker Service

A structure that keeps track of which training jobs launched by your hyperparameter tuning job are not improving model performance as evaluated against an objective function.

Contents

MaxNumberOfTrainingJobsNotImproving

The number of training jobs that have failed to improve model performance by 1% or greater over prior training jobs as evaluated against an objective function.

Type: Integer

Valid Range: Minimum value of 3.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Bias

Service: Amazon SageMaker Service

Contains bias metrics for a model.

Contents

PostTrainingReport

The post-training bias report for a model.

Type: MetricsSource object

Required: No

PreTrainingReport

The pre-training bias report for a model.

Type: MetricsSource object

Required: No

Report

The bias report for a model

Type: MetricsSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BlueGreenUpdatePolicy

Service: Amazon SageMaker Service

Update policy for a blue/green deployment. If this update policy is specified, SageMaker creates a new fleet during the deployment while maintaining the old fleet. SageMaker flips traffic to the new fleet according to the specified traffic routing configuration. Only one update policy should be used in the deployment configuration. If no update policy is specified, SageMaker uses a blue/green deployment strategy with all at once traffic shifting by default.

Contents

TrafficRoutingConfiguration

Defines the traffic routing strategy to shift traffic from the old fleet to the new fleet during an endpoint deployment.

Type: TrafficRoutingConfig object

Required: Yes

MaximumExecutionTimeoutInSeconds

Maximum execution timeout for the deployment. Note that the timeout value should be larger than the total waiting time specified in `TerminationWaitInSeconds` and `WaitIntervalInSeconds`.

Type: Integer

Valid Range: Minimum value of 600. Maximum value of 28800.

Required: No

TerminationWaitInSeconds

Additional waiting time in seconds after the completion of an endpoint deployment before terminating the old endpoint fleet. Default is 0.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 3600.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CacheHitResult

Service: Amazon SageMaker Service

Details on the cache hit of a pipeline execution step.

Contents

SourcePipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CallbackStepMetadata

Service: Amazon SageMaker Service

Metadata about a callback step.

Contents

CallbackToken

The pipeline generated token from the Amazon SQS queue.

Type: String

Length Constraints: Fixed length of 10.

Pattern: `[a-zA-Z0-9]+`

Required: No

OutputParameters

A list of the output parameters of the callback step.

Type: Array of OutputParameter objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

SqsQueueUrl

The URL of the Amazon Simple Queue Service (Amazon SQS) queue used by the callback step.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CandidateArtifactLocations

Service: Amazon SageMaker Service

The location of artifacts for an AutoML candidate job.

Contents

Explainability

The Amazon S3 prefix to the explainability artifacts generated for the AutoML candidate.

Type: String

Length Constraints: Minimum length of 1.

Required: Yes

BacktestResults

The Amazon S3 prefix to the accuracy metrics and the inference results observed over the testing window. Available only for the time-series forecasting problem type.

Type: String

Length Constraints: Minimum length of 1.

Required: No

ModelInsights

The Amazon S3 prefix to the model insight artifacts generated for the AutoML candidate.

Type: String

Length Constraints: Minimum length of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

CandidateGenerationConfig

Service: Amazon SageMaker Service

Stores the configuration information for how model candidates are generated using an AutoML job V2.

Contents

AlgorithmsConfig

Your Autopilot job trains a default set of algorithms on your dataset. For tabular and time-series data, you can customize the algorithm list by selecting a subset of algorithms for your problem type.

`AlgorithmsConfig` stores the customized selection of algorithms to train on your data.

- **For the tabular problem type `TabularJobConfig`,** the list of available algorithms to choose from depends on the training mode set in `AutoMLJobConfig.Mode`.
  - `AlgorithmsConfig` should not be set when the training mode `AutoMLJobConfig.Mode` is set to `AUTO`.
  - When `AlgorithmsConfig` is provided, one `AutoMLAlgorithms` attribute must be set and one only.

    If the list of algorithms provided as values for `AutoMLAlgorithms` is empty, `CandidateGenerationConfig` uses the full set of algorithms for the given training mode.
  - When `AlgorithmsConfig` is not provided, `CandidateGenerationConfig` uses the full set of algorithms for the given training mode.

  For the list of all algorithms per training mode, see AlgorithmConfig.

  For more information on each algorithm, see the Algorithm support section in the Autopilot developer guide.
- **For the time-series forecasting problem type `TimeSeriesForecastingJobConfig`,** choose your algorithms from the list provided in AlgorithmConfig.

  For more information on each algorithm, see the Algorithms support for time-series forecasting section in the Autopilot developer guide.

- When `AlgorithmsConfig` is provided, one `AutoMLAlgorithms` attribute must be set and one only.

  If the list of algorithms provided as values for `AutoMLAlgorithms` is empty, `CandidateGenerationConfig` uses the full set of algorithms for time-series forecasting.
- When `AlgorithmsConfig` is not provided, `CandidateGenerationConfig` uses the full set of algorithms for time-series forecasting.

Type: Array of AutoMLAlgorithmConfig objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CandidateProperties

Service: Amazon SageMaker Service

The properties of an AutoML candidate job.

Contents

CandidateArtifactLocations

The Amazon S3 prefix to the artifacts generated for an AutoML candidate.

Type: CandidateArtifactLocations object

Required: No

CandidateMetrics

Information about the candidate metrics for an AutoML job.

Type: Array of MetricDatum objects

Array Members: Minimum number of 0 items. Maximum number of 40 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CanvasAppSettings

Service: Amazon SageMaker Service

The SageMaker Canvas application settings.

Contents

DirectDeploySettings

The model deployment settings for the SageMaker Canvas application.

Type: DirectDeploySettings object

Required: No

EmrServerlessSettings

The settings for running Amazon EMR Serverless data processing jobs in SageMaker Canvas.

Type: EmrServerlessSettings object

Required: No

GenerativeAiSettings

The generative AI settings for the SageMaker Canvas application.

Type: GenerativeAiSettings object

Required: No

IdentityProviderOAuthSettings

The settings for connecting to an external data source with OAuth.

Type: Array of IdentityProviderOAuthSetting objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

KendraSettings

The settings for document querying.

Type: KendraSettings object

Required: No

ModelRegisterSettings

The model registry settings for the SageMaker Canvas application.

Type: ModelRegisterSettings object

Required: No

TimeSeriesForecastingSettings

Time series forecast settings for the SageMaker Canvas application.

Type: TimeSeriesForecastingSettings object

Required: No

WorkspaceSettings

The workspace settings for the SageMaker Canvas application.

Type: WorkspaceSettings object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CapacityReservation

Service: Amazon SageMaker Service

Information about the Capacity Reservation used by an instance or instance group.

Contents

Arn

The Amazon Resource Name (ARN) of the Capacity Reservation.

Type: String

Required: No

Type

The type of Capacity Reservation. Valid values are `ODCR` (On-Demand Capacity Reservation) or `CRG` (Capacity Reservation Group).

Type: String

Valid Values: `ODCR` | `CRG`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CapacitySize

Service: Amazon SageMaker Service

Specifies the type and size of the endpoint capacity to activate for a blue/green deployment, a rolling deployment, or a rollback strategy. You can specify your batches as either instance count or the overall percentage or your fleet.

For a rollback strategy, if you don't specify the fields in this object, or if you set the `Value` to 100%, then SageMaker uses a blue/green rollback strategy and rolls all traffic back to the blue fleet.

Contents

Type

Specifies the endpoint capacity type.

- `INSTANCE_COUNT`: The endpoint activates based on the number of instances.
- `CAPACITY_PERCENT`: The endpoint activates based on the specified percentage of capacity.

Type: String

Valid Values: `INSTANCE_COUNT | CAPACITY_PERCENT`

Required: Yes

Value

Defines the capacity size, either as a number of instances or a capacity percentage.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

CapacitySizeConfig

Service: Amazon SageMaker Service

The configuration of the size measurements of the AMI update. Using this configuration, you can specify whether SageMaker should update your instance group by an amount or percentage of instances.

Contents

Type

Specifies whether SageMaker should process the update by amount or percentage of instances.

Type: String

Valid Values: `INSTANCE_COUNT | CAPACITY_PERCENTAGE`

Required: Yes

Value

Specifies the amount or percentage of instances SageMaker updates at a time.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CaptureContentTypeHeader

Service: Amazon SageMaker Service

Configuration specifying how to treat different headers. If no headers are specified Amazon SageMaker AI will by default base64 encode when capturing the data.

Contents

CsvContentTypes

The list of all content type headers that Amazon SageMaker AI will treat as CSV and capture accordingly.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*\/[a-zA-Z0-9](-*[a-zA-Z0-9.])*`

Required: No

JsonContentTypes

The list of all content type headers that SageMaker AI will treat as JSON and capture accordingly.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*\/[a-zA-Z0-9](-*[a-zA-Z0-9.])*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CaptureOption

Service: Amazon SageMaker Service

Specifies data Model Monitor will capture.

Contents

CaptureMode

Specify the boundary of data to capture.

Type: String

Valid Values: `Input | Output | InputAndOutput`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CategoricalParameter

Service: Amazon SageMaker Service

Environment parameters you want to benchmark your load test against.

Contents

Name

The Name of the environment variable.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Required: Yes

Value

The list of values you can pass.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 3 items.

Length Constraints: Minimum length of 0. Maximum length of 128.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CategoricalParameterRange

Service: Amazon SageMaker Service

A list of categorical hyperparameters to tune.

Contents

Name

The name of the categorical hyperparameter to tune.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

Values

A list of the categories for the hyperparameter.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 30 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CategoricalParameterRangeSpecification

Service: Amazon SageMaker Service

Defines the possible values for a categorical hyperparameter.

Contents

Values

The allowed categories for the hyperparameter.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 30 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CfnCreateTemplateProvider

Service: Amazon SageMaker Service

The CloudFormation template provider configuration for creating infrastructure resources.

Contents

TemplateName

A unique identifier for the template within the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `(?=.{1,32}$)[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

TemplateURL

The Amazon S3 URL of the CloudFormation template.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `(?=.{1,1024}$)(https)://([^/]+)/(.+)`

Required: Yes

Parameters

An array of CloudFormation stack parameters.

Type: Array of CfnStackCreateParameter objects

Array Members: Minimum number of 0 items. Maximum number of 180 items.

Required: No

RoleARN

The IAM role that CloudFormation assumes when creating the stack.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CfnStackCreateParameter

Service: Amazon SageMaker Service

A key-value pair that represents a parameter for the CloudFormation stack.

Contents

Key

The name of the CloudFormation parameter.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.{1,255}`

Required: Yes

Value

The value of the CloudFormation parameter.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4096.

Pattern: `.{0,4096}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CfnStackDetail

Service: Amazon SageMaker Service

Details about the CloudFormation stack.

Contents

StatusMessage

A human-readable message about the stack's current status.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 4096.

Pattern: `.{1,4096}`

Required: Yes

Id

The unique identifier of the CloudFormation stack.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `(?=.{1,256}$)arn:aws[a-z\-]*:cloudformation:[a-z0-9\-]*:[0-9]{12}:stack/[a-zA-Z][a-zA-Z0-9-]{0,127}/.*`

Required: No

Name

The name of the CloudFormation stack.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[A-Za-z][A-Za-z0-9-]{0,127}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CfnStackParameter

Service: Amazon SageMaker Service

A key-value pair representing a parameter used in the CloudFormation stack.

Contents

Key

The name of the CloudFormation parameter.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.{1,255}`

Required: Yes

Value

The value of the CloudFormation parameter.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4096.

Pattern: `.{0,4096}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CfnStackUpdateParameter

Service: Amazon SageMaker Service

A key-value pair representing a parameter used in the CloudFormation stack.

Contents

Key

The name of the CloudFormation parameter.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.{1,255}`

Required: Yes

Value

The value of the CloudFormation parameter.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4096.

Pattern: `.{0,4096}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CfnTemplateProviderDetail

Service: Amazon SageMaker Service

Details about a CloudFormation template provider configuration and associated provisioning information.

Contents

TemplateName

The unique identifier of the template within the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `(?=.{1,32}$)[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

TemplateURL

The Amazon S3 URL of the CloudFormation template.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `(?=.{1,1024}$)(https)://([^/]+)/(.+)`

Required: Yes

Parameters

An array of CloudFormation stack parameters.

Type: Array of CfnStackParameter objects

Array Members: Minimum number of 0 items. Maximum number of 180 items.

Required: No

RoleARN

The IAM role used by CloudFormation to create the stack.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

StackDetail

Information about the CloudFormation stack created by the template provider.

Type: CfnStackDetail object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CfnUpdateTemplateProvider

Service: Amazon SageMaker Service

Contains configuration details for updating an existing CloudFormation template provider in the project.

Contents

TemplateName

The unique identifier of the template to update within the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `(?=.{1,32}$)[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

TemplateURL

The Amazon S3 URL of the CloudFormation template.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `(?=.{1,1024}$)(https)://([^/]+)/(.+)`

Required: Yes

Parameters

An array of CloudFormation stack parameters.

Type: Array of CfnStackUpdateParameter objects

Array Members: Minimum number of 0 items. Maximum number of 180 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Channel

Service: Amazon SageMaker Service

A channel is a named input source that training algorithms can consume.

Contents

ChannelName

The name of the channel.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[A-Za-z0-9\.\-_]+`

Required: Yes

DataSource

The location of the channel data.

Type: DataSource object

Required: Yes

CompressionType

If training data is compressed, the compression type. The default value is `None`. `CompressionType` is used only in Pipe input mode. In File mode, leave this field unset or set it to None.

Type: String

Valid Values: `None | Gzip`

Required: No

ContentType

The MIME type of the data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

InputMode

(Optional) The input mode to use for the data channel in a training job. If you don't set a value for `InputMode`, SageMaker uses the value set for `TrainingInputMode`. Use this parameter to override the `TrainingInputMode` setting in a AlgorithmSpecification request when you have a channel that needs a different input mode from the training job's general setting. To download the data from Amazon Simple Storage Service (Amazon S3) to the provisioned ML storage volume, and mount the directory to a Docker volume, use `File` input mode. To stream data directly from Amazon S3 to the container, choose `Pipe` input mode.

To use a model for incremental training, choose `File` input model.

Type: String

Valid Values: `Pipe | File | FastFile`

Required: No

RecordWrapperType

Specify RecordIO as the value when input data is in raw format but the training algorithm requires the RecordIO format. In this case, SageMaker wraps each individual S3 object in a RecordIO record. If the input data is already in RecordIO format, you don't need to set this attribute. For more information, see Create a Dataset Using RecordIO.

In File mode, leave this field unset or set it to None.

Type: String

Valid Values: `None | RecordIO`

Required: No

ShuffleConfig

A configuration for a shuffle option for input data in a channel. If you use `S3Prefix` for `S3DataType`, this shuffles the results of the S3 key prefix matches. If you use `ManifestFile`,

the order of the S3 object references in the `ManifestFile` is shuffled. If you use `AugmentedManifestFile`, the order of the JSON lines in the `AugmentedManifestFile` is shuffled. The shuffling order is determined using the `Seed` value.

For Pipe input mode, shuffling is done at the start of every epoch. With large datasets this ensures that the order of the training data is different for each epoch, it helps reduce bias and possible overfitting. In a multi-node training job when ShuffleConfig is combined with `S3DataDistributionType` of `ShardedByS3Key`, the data is shuffled across nodes so that the content sent to a particular node on the first epoch might be sent to a different node on the second epoch.

Type: ShuffleConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ChannelSpecification

Service: Amazon SageMaker Service

Defines a named input source, called a channel, to be used by an algorithm.

Contents

Name

The name of the channel.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[A-Za-z0-9\.\-_]+`

Required: Yes

SupportedContentTypes

The supported MIME types for the data.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

SupportedInputModes

The allowed input mode, either FILE or PIPE.

In FILE mode, Amazon SageMaker copies the data from the input source onto the local Amazon Elastic Block Store (Amazon EBS) volumes before starting your training algorithm. This is the most commonly used input mode.

In PIPE mode, Amazon SageMaker streams input data from the source directly to your algorithm without using the EBS volume.

Type: Array of strings

Array Members: Minimum number of 1 item.

Valid Values: `Pipe | File | FastFile`

Required: Yes

Description

A brief description of the channel.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

IsRequired

Indicates whether the channel is required by the algorithm.

Type: Boolean

Required: No

SupportedCompressionTypes

The allowed compression types, if data compression is used.

Type: Array of strings

Valid Values: `None | Gzip`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CheckpointConfig

Service: Amazon SageMaker Service

Contains information about the output location for managed spot training checkpoint data.

Contents

S3Uri

Identifies the S3 path where you want SageMaker to store checkpoints. For example, `s3://bucket-name/key-name-prefix`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

LocalPath

(Optional) The local directory where checkpoints are written. The default directory is `/opt/ml/checkpoints/`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4096.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClarifyCheckStepMetadata

Service: Amazon SageMaker Service

The container for the metadata for the ClarifyCheck step. For more information, see the topic on ClarifyCheck step in the *Amazon SageMaker Developer Guide*.

Contents

BaselineUsedForDriftCheckConstraints

The Amazon S3 URI of baseline constraints file to be used for the drift check.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

CalculatedBaselineConstraints

The Amazon S3 URI of the newly calculated baseline constraints file.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

CheckJobArn

The Amazon Resource Name (ARN) of the check processing job that was run by this step's execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

CheckType

The type of the Clarify Check step

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

ModelPackageGroupName

The model package group name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

RegisterNewBaseline

This flag indicates if a newly calculated baseline can be accessed through step properties `BaselineUsedForDriftCheckConstraints` and `BaselineUsedForDriftCheckStatistics`. If it is set to `False`, the previous baseline of the configured check type must also be available. These can be accessed through the `BaselineUsedForDriftCheckConstraints` property.

Type: Boolean

Required: No

SkipCheck

This flag indicates if the drift check against the previous baseline will be skipped or not. If it is set to `False`, the previous baseline of the configured check type must be available.

Type: Boolean

Required: No

ViolationReport

The Amazon S3 URI of the violation report if violations are detected.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClarifyExplainerConfig

Service: Amazon SageMaker Service

The configuration parameters for the SageMaker Clarify explainer.

Contents

ShapConfig

The configuration for SHAP analysis.

Type: ClarifyShapConfig object

Required: Yes

EnableExplanations

A JMESPath boolean expression used to filter which records to explain. Explanations are activated by default. See `EnableExplanations`for additional information.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.*`

Required: No

InferenceConfig

The inference configuration parameter for the model container.

Type: ClarifyInferenceConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ClarifyInferenceConfig

Service: Amazon SageMaker Service

The inference configuration parameter for the model container.

Contents

ContentTemplate

A template string used to format a JSON record into an acceptable model container input. For example, a `ContentTemplate` string `'{"myfeatures":$features}'` will format a list of features `[1,2,3]` into the record string `'{"myfeatures":[1,2,3]}'`. Required only when the model container input is in JSON Lines format.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.*`

Required: No

FeatureHeaders

The names of the features. If provided, these are included in the endpoint response payload to help readability of the `InvokeEndpoint` output. See the Response section under **Invoke the endpoint** in the Developer Guide for more information.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 256 items.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.*`

Required: No

FeaturesAttribute

Provides the JMESPath expression to extract the features from a model container input in JSON Lines format. For example, if `FeaturesAttribute` is the JMESPath expression

`'myfeatures'`, it extracts a list of features `[1,2,3]` from request data `'{"myfeatures": [1,2,3]}'`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.*`

Required: No

FeatureTypes

A list of data types of the features (optional). Applicable only to NLP explainability. If provided, `FeatureTypes` must have at least one `'text'` string (for example, `['text']`). If `FeatureTypes` is not provided, the explainer infers the feature types based on the baseline data. The feature types are included in the endpoint response payload. For additional information see the response section under **Invoke the endpoint** in the Developer Guide for more information.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 256 items.

Valid Values: `numerical | categorical | text`

Required: No

LabelAttribute

A JMESPath expression used to locate the list of label headers in the model container output.

Example: If the model container output of a batch request is `'{"labels": ["cat","dog","fish"],"probability":[0.6,0.3,0.1]}'`, then set `LabelAttribute` to `'labels'` to extract the list of label headers `["cat","dog","fish"]`

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.*`

Required: No

LabelHeaders

For multiclass classification problems, the label headers are the names of the classes. Otherwise, the label header is the name of the predicted label. These are used to help readability for the output of the `InvokeEndpoint` API. See the response section under **Invoke the endpoint** in the Developer Guide for more information. If there are no label headers in the model container output, provide them manually using this parameter.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.*`

Required: No

LabelIndex

A zero-based index used to extract a label header or list of label headers from model container output in CSV format.

Example for a multiclass model: If the model container output consists of label headers followed by probabilities: `'"[\'cat\',\'dog\',\'fish\']","[0.1,0.6,0.3]"'`, set `LabelIndex` to `0` to select the label headers `['cat','dog','fish']`.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

MaxPayloadInMB

The maximum payload size (MB) allowed of a request from the explainer to the model container. Defaults to 6 MB.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 25.

Required: No

MaxRecordCount

The maximum number of records in a request that the model container can process when querying the model container for the predictions of a synthetic dataset. A record is a unit of input data that inference can be made on, for example, a single line in CSV data. If `MaxRecordCount` is `1`, the model container expects one record per request. A value of 2 or greater means that the model expects batch requests, which can reduce overhead and speed up the inferencing process. If this parameter is not provided, the explainer will tune the record count per request according to the model container's capacity at runtime.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

ProbabilityAttribute

A JMESPath expression used to extract the probability (or score) from the model container output if the model container is in JSON Lines format.

Example: If the model container output of a single request is `'{"predicted_label":1,"probability":0.6}'`, then set `ProbabilityAttribute` to `'probability'`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.*`

Required: No

ProbabilityIndex

A zero-based index used to extract a probability value (score) or list from model container output in CSV format. If this value is not provided, the entire model container output will be treated as a probability value (score) or list.

Example for a single class model: If the model container output consists of a string-formatted prediction label followed by its probability: `'1,0.6'`, set `ProbabilityIndex` to `1` to select the probability value `0.6`.

Example for a multiclass model: If the model container output consists of a string-formatted prediction label followed by its probability: `'"[\'cat\',\'dog\',\'fish \']","[0.1,0.6,0.3]"'`, set `ProbabilityIndex` to 1 to select the probability values `[0.1,0.6,0.3]`.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClarifyShapBaselineConfig

Service: Amazon SageMaker Service

The configuration for the SHAP baseline (also called the background or reference dataset) of the Kernal SHAP algorithm.

> **Note**
>
> - The number of records in the baseline data determines the size of the synthetic dataset, which has an impact on latency of explainability requests. For more information, see the **Synthetic data** of Configure and create an endpoint.
> - `ShapBaseline` and `ShapBaselineUri` are mutually exclusive parameters. One or the either is required to configure a SHAP baseline.

Contents

MimeType

The MIME type of the baseline data. Choose from `'text/csv'` or `'application/jsonlines'`. Defaults to `'text/csv'`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*\/[a-zA-Z0-9](-*[a-zA-Z0-9+.])*`

Required: No

ShapBaseline

The inline SHAP baseline data in string format. `ShapBaseline` can have one or multiple records to be used as the baseline dataset. The format of the SHAP baseline file should be the same format as the training dataset. For example, if the training dataset is in CSV format and each record contains four features, and all features are numerical, then the format of the baseline data should also share these characteristics. For natural language processing (NLP) of text columns, the baseline value should be the value used to replace the unit of text specified by the `Granularity` of the `TextConfig` parameter. The size limit for `ShapBasline` is 4 KB. Use the `ShapBaselineUri` parameter if you want to provide more than 4 KB of baseline data.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 4096.

Pattern: `[\s\S]+`

Required: No

ShapBaselineUri

The uniform resource identifier (URI) of the S3 bucket where the SHAP baseline file is stored. The format of the SHAP baseline file should be the same format as the format of the training dataset. For example, if the training dataset is in CSV format, and each record in the training dataset has four features, and all features are numerical, then the baseline file should also have this same format. Each record should contain only the features. If you are using a virtual private cloud (VPC), the `ShapBaselineUri` should be accessible to the VPC. For more information about setting up endpoints with Amazon Virtual Private Cloud, see Give SageMaker access to Resources in your Amazon Virtual Private Cloud.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClarifyShapConfig

Service: Amazon SageMaker Service

The configuration for SHAP analysis using SageMaker Clarify Explainer.

Contents

ShapBaselineConfig

The configuration for the SHAP baseline of the Kernal SHAP algorithm.

Type: ClarifyShapBaselineConfig object

Required: Yes

NumberOfSamples

The number of samples to be used for analysis by the Kernal SHAP algorithm.

> **Note**
>
> The number of samples determines the size of the synthetic dataset, which has an impact on latency of explainability requests. For more information, see the **Synthetic data** of Configure and create an endpoint.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

Seed

The starting value used to initialize the random number generator in the explainer. Provide a value for this parameter to obtain a deterministic SHAP result.

Type: Integer

Required: No

TextConfig

A parameter that indicates if text features are treated as text and explanations are provided for individual units of text. Required for natural language processing (NLP) explainability only.

Type: ClarifyTextConfig object

Required: No

UseLogit

A Boolean toggle to indicate if you want to use the logit function (true) or log-odds units (false) for model predictions. Defaults to false.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClarifyTextConfig

Service: Amazon SageMaker Service

A parameter used to configure the SageMaker Clarify explainer to treat text features as text so that explanations are provided for individual units of text. Required only for natural language processing (NLP) explainability.

Contents

Granularity

The unit of granularity for the analysis of text features. For example, if the unit is `'token'`, then each token (like a word in English) of the text is treated as a feature. SHAP values are computed for each unit/feature.

Type: String

Valid Values: `token | sentence | paragraph`

Required: Yes

Language

Specifies the language of the text features in ISO 639-1 or ISO 639-3 code of a supported language.

> **Note**
>
> For a mix of multiple languages, use code `'xx'`.

Type: String

Valid Values: `af | sq | ar | hy | eu | bn | bg | ca | zh | hr | cs | da | nl | en | et | fi | fr | de | el | gu | he | hi | hu | is | id | ga | it | kn | ky | lv | lt | lb | mk | ml | mr | ne | nb | fa | pl | pt | ro | ru | sa | sr | tn | si | sk | sl | es | sv | tl | ta | tt | te | tr | uk | ur | yo | lij | xx`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterAutoScalingConfig

Service: Amazon SageMaker Service

Specifies the autoscaling configuration for a HyperPod cluster.

Contents

Mode

Describes whether autoscaling is enabled or disabled for the cluster. Valid values are `Enable` and `Disable`.

Type: String

Valid Values: `Enable | Disable`

Required: Yes

AutoScalerType

The type of autoscaler to use. Currently supported value is `Karpenter`.

Type: String

Valid Values: `Karpenter`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterAutoScalingConfigOutput

Service: Amazon SageMaker Service

The autoscaling configuration and status information for a HyperPod cluster.

Contents

Mode

Describes whether autoscaling is enabled or disabled for the cluster.

Type: String

Valid Values: `Enable | Disable`

Required: Yes

Status

The current status of the autoscaling configuration. Valid values are `InService`, `Failed`, `Creating`, and `Deleting`.

Type: String

Valid Values: `InService | Failed | Creating | Deleting`

Required: Yes

AutoScalerType

The type of autoscaler configured for the cluster.

Type: String

Valid Values: `Karpenter`

Required: No

FailureMessage

If the autoscaling status is `Failed`, this field contains a message describing the failure.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterCapacityRequirements

Service: Amazon SageMaker Service

Defines the instance capacity requirements for an instance group, including configurations for both Spot and On-Demand capacity types.

Contents

OnDemand

Configuration options specific to On-Demand instances.

Type: ClusterOnDemandOptions object

Required: No

Spot

Configuration options specific to Spot instances.

Type: ClusterSpotOptions object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterEbsVolumeConfig

Service: Amazon SageMaker Service

Defines the configuration for attaching an additional Amazon Elastic Block Store (EBS) volume to each instance of the SageMaker HyperPod cluster instance group. To learn more, see SageMaker HyperPod release notes: June 20, 2024.

Contents

RootVolume

Specifies whether the configuration is for the cluster's root or secondary Amazon EBS volume. You can specify two `ClusterEbsVolumeConfig` fields to configure both the root and secondary volumes. Set the value to `True` if you'd like to provide your own customer managed AWS KMS key to encrypt the root volume. When `True`:

- The configuration is applied to the root volume.
- You can't specify the `VolumeSizeInGB` field. The size of the root volume is determined for you.
- You must specify a KMS key ID for `VolumeKmsKeyId` to encrypt the root volume with your own KMS key instead of an AWS owned KMS key.

Otherwise, by default, the value is `False`, and the following applies:

- The configuration is applied to the secondary volume, while the root volume is encrypted with an AWS owned key.
- You must specify the `VolumeSizeInGB` field.
- You can optionally specify the `VolumeKmsKeyId` to encrypt the secondary volume with your own KMS key instead of an AWS owned KMS key.

Type: Boolean

Required: No

VolumeKmsKeyId

The ID of a KMS key to encrypt the Amazon EBS volume.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

VolumeSizeInGB

The size in gigabytes (GB) of the additional EBS volume to be attached to the instances in the SageMaker HyperPod cluster instance group. The additional EBS volume is attached to each instance within the SageMaker HyperPod cluster instance group and mounted to `/opt/sagemaker`.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 16384.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterEventDetail

Service: Amazon SageMaker Service

Detailed information about a specific event in a HyperPod cluster.

Contents

ClusterArn

The Amazon Resource Name (ARN) of the HyperPod cluster associated with the event.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Required: Yes

ClusterName

The name of the HyperPod cluster associated with the event.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

EventId

The unique identifier (UUID) of the event.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

EventTime

The timestamp when the event occurred.

Type: Timestamp

Required: Yes

ResourceType

The type of resource associated with the event. Valid values are `Cluster`, `InstanceGroup`, or `Instance`.

Type: String

Valid Values: `Cluster | InstanceGroup | Instance`

Required: Yes

Description

A human-readable description of the event.

Type: String

Required: No

EventDetails

Additional details about the event, including event-specific metadata.

Type: EventDetails object

Required: No

EventLevel

The severity level of the event. Valid values are `Info`, `Warn`, and `Error`.

Type: String

Valid Values: `Info | Warn | Error`

Required: No

InstanceGroupName

The name of the instance group associated with the event, if applicable.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

InstanceId

The EC2 instance ID associated with the event, if applicable.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterEventSummary

Service: Amazon SageMaker Service

A summary of an event in a HyperPod cluster.

Contents

ClusterArn

The Amazon Resource Name (ARN) of the HyperPod cluster associated with the event.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Required: Yes

ClusterName

The name of the HyperPod cluster associated with the event.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

EventId

The unique identifier (UUID) of the event.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

EventTime

The timestamp when the event occurred.

Type: Timestamp

Required: Yes

ResourceType

The type of resource associated with the event. Valid values are `Cluster`, `InstanceGroup`, or `Instance`.

Type: String

Valid Values: `Cluster | InstanceGroup | Instance`

Required: Yes

Description

A brief, human-readable description of the event.

Type: String

Required: No

EventLevel

The severity level of the event. Valid values are `Info`, `Warn`, and `Error`.

Type: String

Valid Values: `Info | Warn | Error`

Required: No

InstanceGroupName

The name of the instance group associated with the event, if applicable.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

InstanceId

The Amazon Elastic Compute Cloud (EC2) instance ID associated with the event, if applicable.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterFsxLustreConfig

Service: Amazon SageMaker Service

Defines the configuration for attaching an Amazon FSx for Lustre file system to instances in a SageMaker HyperPod cluster instance group.

Contents

DnsName

The DNS name of the Amazon FSx for Lustre file system.

Type: String

Length Constraints: Minimum length of 16. Maximum length of 275.

Pattern: `((fs|fc)i?-[0-9a-f]{8,}\..{4,253})`

Required: Yes

MountName

The mount name of the Amazon FSx for Lustre file system.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 8.

Pattern: `([A-Za-z0-9_-]{1,8})`

Required: Yes

MountPath

The local path where the Amazon FSx for Lustre file system is mounted on instances.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `/[a-zA-Z0-9._/-]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterFsxOpenZfsConfig

Service: Amazon SageMaker Service

Defines the configuration for attaching an Amazon FSx for OpenZFS file system to instances in a SageMaker HyperPod cluster instance group.

Contents

DnsName

The DNS name of the Amazon FSx for OpenZFS file system.

Type: String

Length Constraints: Minimum length of 16. Maximum length of 275.

Pattern: `((fs|fc)i?-[0-9a-f]{8,}\..{4,253})`

Required: Yes

MountPath

The local path where the Amazon FSx for OpenZFS file system is mounted on instances.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `/[a-zA-Z0-9._/-]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterInstanceGroupDetails

Service: Amazon SageMaker Service

Details of an instance group in a SageMaker HyperPod cluster.

Contents

ActiveOperations

A map indicating active operations currently in progress for the instance group of a SageMaker HyperPod cluster. When there is a scaling operation in progress, this map contains a key `Scaling` with value 1.

Type: String to integer map

Valid Keys: `Scaling`

Valid Range: Minimum value of 1.

Required: No

ActiveSoftwareUpdateConfig

The configuration to use when updating the AMI versions.

Type: DeploymentConfiguration object

Required: No

CapacityRequirements

The instance capacity requirements for the instance group.

Type: ClusterCapacityRequirements object

Required: No

CurrentCount

The number of instances that are currently in the instance group of a SageMaker HyperPod cluster.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

CurrentImageId

The ID of the Amazon Machine Image (AMI) currently in use by the instance group.

Type: String

Length Constraints: Minimum length of 7. Maximum length of 21.

Pattern: `ami-[0-9a-fA-F]{8,17}|default`

Required: No

DesiredImageId

The ID of the Amazon Machine Image (AMI) desired for the instance group.

Type: String

Length Constraints: Minimum length of 7. Maximum length of 21.

Pattern: `ami-[0-9a-fA-F]{8,17}|default`

Required: No

ExecutionRole

The execution role for the instance group to assume.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

ImageVersionStatus

The status of the image version for the instance group. Indicates whether the instance group is running the latest image version or if an update is available.

Type: String

Valid Values: `UpToDate | UpdateAvailable`

Required: No

InstanceGroupName

The name of the instance group of a SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

InstanceRequirements

The instance requirements for the instance group, including the current and desired instance types. This field is present for flexible instance groups that support multiple instance types.

Type: ClusterInstanceRequirementDetails object

Required: No

InstanceStorageConfigs

The additional storage configurations for the instances in the SageMaker HyperPod cluster instance group.

Type: Array of ClusterInstanceStorageConfig objects

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Required: No

InstanceType

The instance type of the instance group of a SageMaker HyperPod cluster.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge |`

```
ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge
| ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge
| ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge
| ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge
| ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large
| ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge |
ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge |
ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge |
ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge |
ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge
| ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge |
ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge |
ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge
| ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large
| ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge |
ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge
| ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge |
ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge
| ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge |
ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge
| ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge |
ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge
| ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge |
ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge |
ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge |
ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge |
ml.g7e.48xlarge | ml.p6-b300.48xlarge
```

Required: No

InstanceTypeDetails

Details about the instance types in the instance group, including the count and configuration of each instance type. This field is present for flexible instance groups that support multiple instance types.

Type: Array of ClusterInstanceTypeDetail objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

KubernetesConfig

The Kubernetes configuration for the instance group that contains labels and taints to be applied for the nodes in this instance group.

Type: ClusterKubernetesConfigDetails object

Required: No

LifeCycleConfig

Details of LifeCycle configuration for the instance group.

Type: ClusterLifeCycleConfig object

Required: No

MinCount

The minimum number of instances that must be available in the instance group of a SageMaker HyperPod cluster before it transitions to `InService` status.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 6758.

Required: No

NetworkInterface

The network interface configuration for the instance group.

Type: ClusterNetworkInterfaceDetails object

Required: No

OnStartDeepHealthChecks

A flag indicating whether deep health checks should be performed when the cluster instance group is created or updated.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `InstanceStress | InstanceConnectivity`

Required: No

OverrideVpcConfig

The customized Amazon VPC configuration at the instance group level that overrides the default Amazon VPC configuration of the SageMaker HyperPod cluster.

Type: VpcConfig object

Required: No

ScheduledUpdateConfig

The configuration object of the schedule that SageMaker follows when updating the AMI.

Type: ScheduledUpdateConfig object

Required: No

SlurmConfig

The Slurm configuration for the instance group.

Type: ClusterSlurmConfigDetails object

Required: No

SoftwareUpdateStatus

Status of the last software udpate request.

Status transitions follow these possible sequences:

- Pending -> InProgress -> Succeeded
- Pending -> InProgress -> RollbackInProgress -> RollbackComplete
- Pending -> InProgress -> RollbackInProgress -> Failed

Type: String

Valid Values: `Pending | InProgress | Succeeded | Failed | RollbackInProgress | RollbackComplete`

Required: No

Status

The current status of the cluster instance group.

- `InService`: The instance group is active and healthy.
- `Creating`: The instance group is being provisioned.
- `Updating`: The instance group is being updated.
- `Failed`: The instance group has failed to provision or is no longer healthy.
- `Degraded`: The instance group is degraded, meaning that some instances have failed to provision or are no longer healthy.
- `Deleting`: The instance group is being deleted.

Type: String

Valid Values: `InService | Creating | Updating | Failed | Degraded | SystemUpdating | Deleting`

Required: No

TargetCount

The number of instances you specified to add to the instance group of a SageMaker HyperPod cluster.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 6758.

Required: No

TargetStateCount

Represents the number of running nodes using the desired Image ID.

1. **During software update operations:** This count shows the number of nodes running on the desired Image ID. If a rollback occurs, the current image ID and desired image ID (both included in the describe cluster response) swap values. The TargetStateCount then shows the number of nodes running on the newly designated desired image ID (which was previously the current image ID).
2. **During simultaneous scaling and software update operations:** This count shows the number of instances running on the desired image ID, including any new instances created as part of the scaling request. New nodes are always created using the desired image ID, so

TargetStateCount reflects the total count of nodes running on the desired image ID, even during rollback scenarios.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 6758.

Required: No

ThreadsPerCore

The number you specified to `TreadsPerCore` in `CreateCluster` for enabling or disabling multithreading. For instance types that support multithreading, you can specify 1 for disabling multithreading and 2 for enabling multithreading. For more information, see the reference table of CPU cores and threads per CPU core per instance type in the *Amazon Elastic Compute Cloud User Guide*.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 2.

Required: No

TrainingPlanArn

The Amazon Resource Name (ARN); of the training plan associated with this cluster instance group.

For more information about how to reserve GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Required: No

TrainingPlanStatus

The current status of the training plan associated with this cluster instance group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterInstanceGroupSpecification

Service: Amazon SageMaker Service

The specifications of an instance group that you need to define.

Contents

ExecutionRole

Specifies an IAM execution role to be assumed by the instance group.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

InstanceCount

Specifies the number of instances to add to the instance group of a SageMaker HyperPod cluster.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 6758.

Required: Yes

InstanceGroupName

Specifies the name of the instance group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

CapacityRequirements

Specifies the capacity requirements for the instance group.

Type: ClusterCapacityRequirements object

Required: No

ImageId

When configuring your HyperPod cluster, you can specify an image ID using one of the following options:

- `HyperPodPublicAmiId`: Use a HyperPod public AMI
- `CustomAmiId`: Use your custom AMI
- `default`: Use the default latest system image

If you choose to use a custom AMI (`CustomAmiId`), ensure it meets the following requirements:

- Encryption: The custom AMI must be unencrypted.
- Ownership: The custom AMI must be owned by the same AWS account that is creating the HyperPod cluster.
- Volume support: Only the primary AMI snapshot volume is supported; additional AMI volumes are not supported.

When updating the instance group's AMI through the `UpdateClusterSoftware` operation, if an instance group uses a custom AMI, you must provide an `ImageId` or use the default as input. Note that if you don't specify an instance group in your `UpdateClusterSoftware` request, then all of the instance groups are patched with the specified image.

Type: String

Length Constraints: Minimum length of 7. Maximum length of 21.

Pattern: `ami-[0-9a-fA-F]{8,17}|default`

Required: No

InstanceRequirements

The instance requirements for the instance group, including the instance types to use. Use this to create a flexible instance group that supports multiple instance types. The `InstanceType` and `InstanceRequirements` properties are mutually exclusive.

Type: ClusterInstanceRequirements object

Required: No

InstanceStorageConfigs

Specifies the additional storage configurations for the instances in the SageMaker HyperPod cluster instance group.

Type: Array of ClusterInstanceStorageConfig objects

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Required: No

InstanceType

Specifies the instance type of the instance group.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge`

`| ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge | ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge`

Required: No

KubernetesConfig

Specifies the Kubernetes configuration for the instance group. You describe what you want the labels and taints to look like, and the cluster works to reconcile the actual state with the declared state for nodes in this instance group.

Type: ClusterKubernetesConfig object

Required: No

LifeCycleConfig

Specifies the LifeCycle configuration for the instance group.

Type: ClusterLifeCycleConfig object

Required: No

MinInstanceCount

Defines the minimum number of instances required for an instance group to become `InService`. If this threshold isn't met within 3 hours, the instance group rolls back to its previous state - zero instances for new instance groups, or previous settings for existing instance groups. `MinInstanceCount` only affects the initial transition to `InService` and does not guarantee maintaining this minimum afterward.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 6758.

Required: No

NetworkInterface

The network interface configuration for the instance group.

Type: ClusterNetworkInterface object

Required: No

OnStartDeepHealthChecks

A flag indicating whether deep health checks should be performed when the cluster instance group is created or updated.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `InstanceStress | InstanceConnectivity`

Required: No

OverrideVpcConfig

To configure multi-AZ deployments, customize the Amazon VPC configuration at the instance group level. You can specify different subnets and security groups across different AZs in the instance group specification to override a SageMaker HyperPod cluster's default Amazon VPC configuration. For more information about deploying a cluster in multiple AZs, see Setting up SageMaker HyperPod clusters across multiple AZs.

> **Note**
>
> When your Amazon VPC and subnets support IPv6, network communications differ based on the cluster orchestration platform:
>
> - Slurm-orchestrated clusters automatically configure nodes with dual IPv6 and IPv4 addresses, allowing immediate IPv6 network communications.
> - In Amazon EKS-orchestrated clusters, nodes receive dual-stack addressing, but pods can only use IPv6 when the Amazon EKS cluster is explicitly IPv6-enabled. For information about deploying an IPv6 Amazon EKS cluster, see Amazon EKS IPv6 Cluster Deployment.
>
> Additional resources for IPv6 configuration:
>
> - For information about adding IPv6 support to your VPC, see to IPv6 Support for VPC.

- For information about creating a new IPv6-compatible VPC, see Amazon VPC Creation Guide.
- To configure SageMaker HyperPod with a custom Amazon VPC, see Custom Amazon VPC Setup for SageMaker HyperPod.

Type: VpcConfig object

Required: No

ScheduledUpdateConfig

The configuration object of the schedule that SageMaker uses to update the AMI.

Type: ScheduledUpdateConfig object

Required: No

SlurmConfig

Specifies the Slurm configuration for the instance group.

Type: ClusterSlurmConfig object

Required: No

ThreadsPerCore

Specifies the value for **Threads per core**. For instance types that support multithreading, you can specify 1 for disabling multithreading and 2 for enabling multithreading. For instance types that doesn't support multithreading, specify 1. For more information, see the reference table of CPU cores and threads per CPU core per instance type in the *Amazon Elastic Compute Cloud User Guide*.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 2.

Required: No

TrainingPlanArn

The Amazon Resource Name (ARN); of the training plan to use for this cluster instance group.

For more information about how to reserve GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterInstancePlacement

Service: Amazon SageMaker Service

Specifies the placement details for the node in the SageMaker HyperPod cluster, including the Availability Zone and the unique identifier (ID) of the Availability Zone.

Contents

AvailabilityZone

The Availability Zone where the node in the SageMaker HyperPod cluster is launched.

Type: String

Pattern: `[a-z]{2}-[a-z]+-\d[a-z]`

Required: No

AvailabilityZoneId

The unique identifier (ID) of the Availability Zone where the node in the SageMaker HyperPod cluster is launched.

Type: String

Pattern: `[a-z]{3}\d-az\d`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterInstanceRequirementDetails

Service: Amazon SageMaker Service

The instance requirement details for a flexible instance group, including the current and desired instance types.

Contents

CurrentInstanceTypes

The instance types currently in use by the instance group.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge`

| ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge | ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge

Required: No

DesiredInstanceTypes

The desired instance types for the instance group, as specified in the most recent update request.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Valid Values: ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge

```
| ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large
| ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge |
ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge
| ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge |
ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge
| ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge |
ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge
| ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge |
ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge
| ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge |
ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge |
ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge |
ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge |
ml.g7e.48xlarge | ml.p6-b300.48xlarge
```

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterInstanceRequirements

Service: Amazon SageMaker Service

The instance requirements for a flexible instance group. Use this to specify multiple instance types that the instance group can use. The order of instance types in the list determines the priority for instance provisioning.

Contents

InstanceTypes

The list of instance types that the instance group can use. The order of instance types determines the priority—HyperPod attempts to provision instances using the first instance type in the list and falls back to subsequent types if capacity is unavailable.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge |`

`ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge | ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterInstanceStatusDetails

Service: Amazon SageMaker Service

Details of an instance in a SageMaker HyperPod cluster.

Contents

Status

The status of an instance in a SageMaker HyperPod cluster.

Type: String

Valid Values: `Running | Failure | Pending | ShuttingDown | SystemUpdating | DeepHealthCheckInProgress | NotFound`

Required: Yes

Message

The message from an instance in a SageMaker HyperPod cluster.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterInstanceStorageConfig

Service: Amazon SageMaker Service

Defines the configuration for attaching additional storage to the instances in the SageMaker HyperPod cluster instance group. To learn more, see SageMaker HyperPod release notes: June 20, 2024.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

EbsVolumeConfig

Defines the configuration for attaching additional Amazon Elastic Block Store (EBS) volumes to the instances in the SageMaker HyperPod cluster instance group. The additional EBS volume is attached to each instance within the SageMaker HyperPod cluster instance group and mounted to `/opt/sagemaker`.

Type: ClusterEbsVolumeConfig object

Required: No

FsxLustreConfig

Defines the configuration for attaching an Amazon FSx for Lustre file system to the instances in the SageMaker HyperPod cluster instance group.

Type: ClusterFsxLustreConfig object

Required: No

FsxOpenZfsConfig

Defines the configuration for attaching an Amazon FSx for OpenZFS file system to the instances in the SageMaker HyperPod cluster instance group.

Type: ClusterFsxOpenZfsConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterInstanceTypeDetail

Service: Amazon SageMaker Service

Details about a specific instance type within a flexible instance group, including the count and configuration.

Contents

CurrentCount

The number of instances of this type currently running in the instance group.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

InstanceType

The instance type.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge |`

ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge | ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge

Required: No

ThreadsPerCore

The number of threads per CPU core for this instance type.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 2.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterKubernetesConfig

Service: Amazon SageMaker Service

Kubernetes configuration that specifies labels and taints to be applied to cluster nodes in an instance group.

Contents

Labels

Key-value pairs of labels to be applied to cluster nodes.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 317.

Key Pattern: `([a-z0-9]([-a-z0-9]*[a-z0-9])?(\.[a-z0-9]([-a-z0-9]*[a-z0-9])?)*/)?[A-Za-z0-9]([-A-Za-z0-9_.]*[A-Za-z0-9])?`

Value Length Constraints: Minimum length of 1. Maximum length of 63.

Value Pattern: `(([A-Za-z0-9][-A-Za-z0-9_.]*)?[A-Za-z0-9])?`

Required: No

Taints

List of taints to be applied to cluster nodes.

Type: Array of ClusterKubernetesTaint objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterKubernetesConfigDetails

Service: Amazon SageMaker Service

Detailed Kubernetes configuration showing both the current and desired state of labels and taints for cluster nodes.

Contents

CurrentLabels

The current labels applied to cluster nodes of an instance group.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 317.

Key Pattern: `([a-z0-9]([-a-z0-9]*[a-z0-9])?(\.[a-z0-9]([-a-z0-9]*[a-z0-9])?)*/)?[A-Za-z0-9]([-A-Za-z0-9_.]*[A-Za-z0-9])?`

Value Length Constraints: Minimum length of 1. Maximum length of 63.

Value Pattern: `(([A-Za-z0-9][-A-Za-z0-9_.]*)?[A-Za-z0-9])?`

Required: No

CurrentTaints

The current taints applied to cluster nodes of an instance group.

Type: Array of ClusterKubernetesTaint objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

DesiredLabels

The desired labels to be applied to cluster nodes of an instance group.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 317.

Key Pattern: `([a-z0-9]([-a-z0-9]*[a-z0-9])?(\.[a-z0-9]([-a-z0-9]*[a-z0-9])?)*/)?[A-Za-z0-9]([-A-Za-z0-9_.]*[A-Za-z0-9])?`

Value Length Constraints: Minimum length of 1. Maximum length of 63.

Value Pattern: `(([A-Za-z0-9][-A-Za-z0-9_.]*)?[A-Za-z0-9])?`

Required: No

DesiredTaints

The desired taints to be applied to cluster nodes of an instance group.

Type: Array of ClusterKubernetesTaint objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterKubernetesConfigNodeDetails

Service: Amazon SageMaker Service

Node-specific Kubernetes configuration showing both current and desired state of labels and taints for an individual cluster node.

Contents

CurrentLabels

The current labels applied to the cluster node.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 317.

Key Pattern: `([a-z0-9]([-a-z0-9]*[a-z0-9])?(\.[a-z0-9]([-a-z0-9]*[a-z0-9])?)*/)?[A-Za-z0-9]([-A-Za-z0-9_.]*[A-Za-z0-9])?`

Value Length Constraints: Minimum length of 1. Maximum length of 63.

Value Pattern: `(([A-Za-z0-9][-A-Za-z0-9_.]*)?[A-Za-z0-9])?`

Required: No

CurrentTaints

The current taints applied to the cluster node.

Type: Array of ClusterKubernetesTaint objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

DesiredLabels

The desired labels to be applied to the cluster node.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 317.

Key Pattern: `([a-z0-9]([-a-z0-9]*[a-z0-9])?(\.[a-z0-9]([-a-z0-9]*[a-z0-9])?)*/)?[A-Za-z0-9]([-A-Za-z0-9_.]*[A-Za-z0-9])?`

Value Length Constraints: Minimum length of 1. Maximum length of 63.

Value Pattern: `(([A-Za-z0-9][-A-Za-z0-9_.]*)?[A-Za-z0-9])?`

Required: No

DesiredTaints

The desired taints to be applied to the cluster node.

Type: Array of ClusterKubernetesTaint objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterKubernetesTaint

Service: Amazon SageMaker Service

A Kubernetes taint that can be applied to cluster nodes.

Contents

Effect

The effect of the taint. Valid values are `NoSchedule`, `PreferNoSchedule`, and `NoExecute`.

Type: String

Valid Values: `NoSchedule | PreferNoSchedule | NoExecute`

Required: Yes

Key

The key of the taint.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 317.

Pattern: `([a-z0-9]([-a-z0-9]*[a-z0-9])?(\.[a-z0-9]([-a-z0-9]*[a-z0-9])?)*/)?[A-Za-z0-9]([-A-Za-z0-9_.]*[A-Za-z0-9])?`

Required: Yes

Value

The value of the taint.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `(([A-Za-z0-9][-A-Za-z0-9_.]*)?[A-Za-z0-9])?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterLifeCycleConfig

Service: Amazon SageMaker Service

The lifecycle configuration for a SageMaker HyperPod cluster.

Contents

OnCreate

The file name of the entrypoint script of lifecycle scripts under `SourceS3Uri`. This entrypoint script runs during cluster creation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\S\s]+`

Required: No

OnInitComplete

The file name of the entrypoint script of lifecycle scripts under `SourceS3Uri`. This script runs on the node after the AMI-based initialization is complete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\S\s]+`

Required: No

SourceS3Uri

An Amazon S3 bucket path where your lifecycle scripts are stored.

> **Important**
>
> Make sure that the S3 bucket path starts with `s3://sagemaker-`.
> The IAM role for SageMaker HyperPod has the managed `AmazonSageMakerClusterInstanceRolePolicy` attached, which allows access to S3 buckets with the specific prefix `sagemaker-`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterMetadata

Service: Amazon SageMaker Service

Metadata information about a HyperPod cluster showing information about the cluster level operations, such as creating, updating, and deleting.

Contents

EksRoleAccessEntries

A list of Amazon EKS IAM role ARNs associated with the cluster. This is created by HyperPod on your behalf and only applies for EKS orchestrated clusters.

Type: Array of strings

Required: No

FailureMessage

An error message describing why the cluster level operation (such as creating, updating, or deleting) failed.

Type: String

Required: No

SlrAccessEntry

The Service-Linked Role (SLR) associated with the cluster. This is created by HyperPod on your behalf and only applies for EKS orchestrated clusters.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ClusterNetworkInterface

Service: Amazon SageMaker Service

The network interface configuration for a Amazon SageMaker HyperPod cluster instance group.

Contents

InterfaceType

The type of network interface for the instance group. Valid values:

- `efa` – An EFA with ENA interface, which provides both the EFA device for low-latency, high-throughput communication and the ENA device for IP networking.
- `efa-only` – An EFA-only interface, which provides only the EFA device capabilities without the ENA device for traditional IP networking.

For more information, see Elastic Fabric Adapter.

Type: String

Valid Values: `efa | efa-only`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterNetworkInterfaceDetails

Service: Amazon SageMaker Service

The network interface configuration details for a Amazon SageMaker HyperPod cluster instance group.

Contents

InterfaceType

The type of network interface for the instance group. Valid values are `efa` and `efa-only`.

Type: String

Valid Values: `efa | efa-only`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterNodeDetails

Service: Amazon SageMaker Service

Details of an instance (also called a *node* interchangeably) in a SageMaker HyperPod cluster.

Contents

CapacityType

The capacity type of the node. Valid values are `OnDemand` and `Spot`. When set to `OnDemand`, the node is launched as an On-Demand instance. When set to `Spot`, the node is launched as a Spot instance.

Type: String

Valid Values: `Spot | OnDemand`

Required: No

CurrentImageId

The ID of the Amazon Machine Image (AMI) currently in use by the node.

Type: String

Length Constraints: Minimum length of 7. Maximum length of 21.

Pattern: `ami-[0-9a-fA-F]{8,17}|default`

Required: No

DesiredImageId

The ID of the Amazon Machine Image (AMI) desired for the node.

Type: String

Length Constraints: Minimum length of 7. Maximum length of 21.

Pattern: `ami-[0-9a-fA-F]{8,17}|default`

Required: No

ImageVersionStatus

The status of the image version for the cluster node.

Type: String

Valid Values: `UpToDate | UpdateAvailable`

Required: No

InstanceGroupName

The instance group name in which the instance is.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

InstanceId

The ID of the instance.

Type: String

Required: No

InstanceStatus

The status of the instance.

Type: ClusterInstanceStatusDetails object

Required: No

InstanceStorageConfigs

The configurations of additional storage specified to the instance group where the instance (node) is launched.

Type: Array of ClusterInstanceStorageConfig objects

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Required: No

InstanceType

The type of the instance.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge | ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge |`

`ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge`

Required: No

KubernetesConfig

The Kubernetes configuration applied to this node, showing both the current and desired state of labels and taints. The cluster works to reconcile the actual state with the declared state.

Type: ClusterKubernetesConfigNodeDetails object

Required: No

LastSoftwareUpdateTime

The time when the cluster was last updated.

Type: Timestamp

Required: No

LaunchTime

The time when the instance is launched.

Type: Timestamp

Required: No

LifeCycleConfig

The LifeCycle configuration applied to the instance.

Type: ClusterLifeCycleConfig object

Required: No

NetworkInterface

The network interface configuration for the cluster node.

Type: ClusterNetworkInterfaceDetails object

Required: No

NodeLogicalId

A unique identifier for the node that persists throughout its lifecycle, from provisioning request to termination. This identifier can be used to track the node even before it has an assigned `InstanceId`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Required: No

OverrideVpcConfig

The customized Amazon VPC configuration at the instance group level that overrides the default Amazon VPC configuration of the SageMaker HyperPod cluster.

Type: VpcConfig object

Required: No

Placement

The placement details of the SageMaker HyperPod cluster node.

Type: ClusterInstancePlacement object

Required: No

PrivateDnsHostname

The private DNS hostname of the SageMaker HyperPod cluster node.

Type: String

Pattern: `ip-((25[0-5]|(2[0-4]|1\d|[1-9]|)\d)-?\b){4}\..*`

Required: No

PrivatePrimaryIp

The private primary IP address of the SageMaker HyperPod cluster node.

Type: String

Pattern: `((25[0-5]|(2[0-4]|1\d|[1-9]|)\d)\.?\b){4}`

Required: No

PrivatePrimaryIpv6

The private primary IPv6 address of the SageMaker HyperPod cluster node when configured with an Amazon VPC that supports IPv6 and includes subnets with IPv6 addressing enabled in either the cluster Amazon VPC configuration or the instance group Amazon VPC configuration.

Type: String

Required: No

ThreadsPerCore

The number of threads per CPU core you specified under `CreateCluster`.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 2.

Required: No

UltraServerInfo

Contains information about the UltraServer.

Type: UltraServerInfo object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterNodeSummary

Service: Amazon SageMaker Service

Lists a summary of the properties of an instance (also called a *node* interchangeably) of a SageMaker HyperPod cluster.

Contents

InstanceGroupName

The name of the instance group in which the instance is.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

InstanceId

The ID of the instance.

Type: String

Required: Yes

InstanceStatus

The status of the instance.

Type: ClusterInstanceStatusDetails object

Required: Yes

InstanceType

The type of the instance.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge`

| ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge | ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge

Required: Yes

LaunchTime

The time when the instance is launched.

Type: Timestamp

Required: Yes

ImageVersionStatus

The status of the image version for the cluster node.

Type: String

Valid Values: `UpToDate | UpdateAvailable`

Required: No

LastSoftwareUpdateTime

The time when SageMaker last updated the software of the instances in the cluster.

Type: Timestamp

Required: No

NodeLogicalId

A unique identifier for the node that persists throughout its lifecycle, from provisioning request to termination. This identifier can be used to track the node even before it has an assigned `InstanceId`. This field is only included when `IncludeNodeLogicalIds` is set to `True` in the `ListClusterNodes` request.

Type: String

Required: No

PrivateDnsHostname

The private DNS hostname of the SageMaker HyperPod cluster node.

Type: String

Pattern: `ip-((25[0-5]|(2[0-4]|1\d|[1-9]|)\d)-?\b){4}\..*`

Required: No

UltraServerInfo

Contains information about the UltraServer.

Type: UltraServerInfo object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterOnDemandOptions

Service: Amazon SageMaker Service

Configuration options specific to On-Demand instances.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterOrchestrator

Service: Amazon SageMaker Service

The type of orchestrator used for the SageMaker HyperPod cluster.

Contents

Eks

The Amazon EKS cluster used as the orchestrator for the SageMaker HyperPod cluster.

Type: ClusterOrchestratorEksConfig object

Required: No

Slurm

The Slurm orchestrator configuration for the SageMaker HyperPod cluster.

Type: ClusterOrchestratorSlurmConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterOrchestratorEksConfig

Service: Amazon SageMaker Service

The configuration settings for the Amazon EKS cluster used as the orchestrator for the SageMaker HyperPod cluster.

Contents

ClusterArn

The Amazon Resource Name (ARN) of the Amazon EKS cluster associated with the SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:eks:[a-z0-9\-]*:[0-9]{12}:cluster\/[0-9A-Za-z][A-Za-z0-9\-_]{0,99}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterOrchestratorSlurmConfig

Service: Amazon SageMaker Service

The configuration settings for the Slurm orchestrator used with the SageMaker HyperPod cluster.

Contents

SlurmConfigStrategy

The strategy for managing partitions for the Slurm configuration. Valid values are `Managed`, `Overwrite`, and `Merge`.

Type: String

Valid Values: `Overwrite | Managed | Merge`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterRestrictedInstanceGroupDetails

Service: Amazon SageMaker Service

The instance group details of the restricted instance group (RIG).

Contents

CurrentCount

The number of instances that are currently in the restricted instance group of a SageMaker HyperPod cluster.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

EnvironmentConfig

The configuration for the restricted instance groups (RIG) environment.

Type: EnvironmentConfigDetails object

Required: No

ExecutionRole

The execution role for the restricted instance group to assume.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

InstanceGroupName

The name of the restricted instance group of a SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

InstanceStorageConfigs

The additional storage configurations for the instances in the SageMaker HyperPod cluster restricted instance group.

Type: Array of ClusterInstanceStorageConfig objects

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Required: No

InstanceType

The instance type of the restricted instance group of a SageMaker HyperPod cluster.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge |`

`ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge | ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge`

Required: No

OnStartDeepHealthChecks

A flag indicating whether deep health checks should be performed when the cluster's restricted instance group is created or updated.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `InstanceStress | InstanceConnectivity`

Required: No

OverrideVpcConfig

Specifies an Amazon Virtual Private Cloud (VPC) that your SageMaker jobs, hosted models, and compute resources have access to. You can control access to and from your resources by configuring a VPC. For more information, see Give SageMaker Access to Resources in your Amazon VPC.

Type: VpcConfig object

Required: No

ScheduledUpdateConfig

The configuration object of the schedule that SageMaker follows when updating the AMI.

Type: ScheduledUpdateConfig object

Required: No

Status

The current status of the cluster's restricted instance group.

- `InService`: The restricted instance group is active and healthy.
- `Creating`: The restricted instance group is being provisioned.
- `Updating`: The restricted instance group is being updated.
- `Failed`: The restricted instance group has failed to provision or is no longer healthy.
- `Degraded`: The restricted instance group is degraded, meaning that some instances have failed to provision or are no longer healthy.
- `Deleting`: The restricted instance group is being deleted.

Type: String

Valid Values: `InService | Creating | Updating | Failed | Degraded | SystemUpdating | Deleting`

Required: No

TargetCount

The number of instances you specified to add to the restricted instance group of a SageMaker HyperPod cluster.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 6758.

Required: No

ThreadsPerCore

The number you specified to `TreadsPerCore` in `CreateCluster` for enabling or disabling multithreading. For instance types that support multithreading, you can specify 1 for disabling multithreading and 2 for enabling multithreading. For more information, see the reference table of CPU cores and threads per CPU core per instance type in the *Amazon Elastic Compute Cloud User Guide*.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 2.

Required: No

TrainingPlanArn

The Amazon Resource Name (ARN) of the training plan to filter clusters by. For more information about reserving GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Required: No

TrainingPlanStatus

The current status of the training plan associated with this cluster restricted instance group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterRestrictedInstanceGroupsConfig

Service: Amazon SageMaker Service

The configuration for the restricted instance groups (RIG) in the SageMaker HyperPod cluster.

Contents

SharedEnvironmentConfig

The shared environment configuration for the restricted instance groups (RIG).

Type: ClusterSharedEnvironmentConfig object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterRestrictedInstanceGroupsConfigOutput

Service: Amazon SageMaker Service

The output configuration for the restricted instance groups (RIG) in the SageMaker HyperPod cluster.

Contents

SharedEnvironmentConfig

The shared environment configuration details for the restricted instance groups (RIG).

Type: ClusterSharedEnvironmentConfigDetails object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterRestrictedInstanceGroupSpecification

Service: Amazon SageMaker Service

The specifications of a restricted instance group that you need to define.

Contents

ExecutionRole

Specifies an IAM execution role to be assumed by the restricted instance group.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

InstanceCount

Specifies the number of instances to add to the restricted instance group of a SageMaker HyperPod cluster.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 6758.

Required: Yes

InstanceGroupName

Specifies the name of the restricted instance group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

InstanceType

Specifies the instance type of the restricted instance group.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge | ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge`

Required: Yes

EnvironmentConfig

The configuration for the restricted instance groups (RIG) environment.

Type: EnvironmentConfig object

Required: No

InstanceStorageConfigs

Specifies the additional storage configurations for the instances in the SageMaker HyperPod cluster restricted instance group.

Type: Array of ClusterInstanceStorageConfig objects

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Required: No

OnStartDeepHealthChecks

A flag indicating whether deep health checks should be performed when the cluster restricted instance group is created or updated.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `InstanceStress | InstanceConnectivity`

Required: No

OverrideVpcConfig

Specifies an Amazon Virtual Private Cloud (VPC) that your SageMaker jobs, hosted models, and compute resources have access to. You can control access to and from your resources by configuring a VPC. For more information, see Give SageMaker Access to Resources in your Amazon VPC.

Type: VpcConfig object

Required: No

ScheduledUpdateConfig

The configuration object of the schedule that SageMaker follows when updating the AMI.

Type: ScheduledUpdateConfig object

Required: No

ThreadsPerCore

The number you specified to `TreadsPerCore` in `CreateCluster` for enabling or disabling multithreading. For instance types that support multithreading, you can specify 1 for disabling multithreading and 2 for enabling multithreading. For more information, see the reference table of CPU cores and threads per CPU core per instance type in the *Amazon Elastic Compute Cloud User Guide*.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 2.

Required: No

TrainingPlanArn

The Amazon Resource Name (ARN) of the training plan to filter clusters by. For more information about reserving GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterSchedulerConfigSummary

Service: Amazon SageMaker Service

Summary of the cluster policy.

Contents

ClusterSchedulerConfigArn

ARN of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:cluster-scheduler-config/[a-z0-9]{12}`

Required: Yes

ClusterSchedulerConfigId

ID of the cluster policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[a-z0-9]{12}`

Required: Yes

CreationTime

Creation time of the cluster policy.

Type: Timestamp

Required: Yes

Name

Name of the cluster policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Status

Status of the cluster policy.

Type: String

Valid Values: `Creating | CreateFailed | CreateRollbackFailed | Created | Updating | UpdateFailed | UpdateRollbackFailed | Updated | Deleting | DeleteFailed | DeleteRollbackFailed | Deleted`

Required: Yes

ClusterArn

ARN of the cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Required: No

ClusterSchedulerConfigVersion

Version of the cluster policy.

Type: Integer

Required: No

LastModifiedTime

Last modified time of the cluster policy.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterSharedEnvironmentConfig

Service: Amazon SageMaker Service

The shared environment configuration for the restricted instance groups (RIG).

Contents

FSxLustreConfig

Configuration settings for an Amazon FSx for Lustre file system in the shared environment.

Type: FSxLustreConfig object

Required: Yes

FSxLustreDeletionPolicy

The deletion policy for the Amazon FSx for Lustre file system in the shared environment.

Type: String

Valid Values: `DeleteIfNotUsed | Keep`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterSharedEnvironmentConfigDetails

Service: Amazon SageMaker Service

The shared environment configuration details for the restricted instance groups (RIG).

Contents

CurrentFSxLustreConfig

The current Amazon FSx for Lustre file system configuration in the shared environment.

Type: FSxLustreConfig object

Required: No

CurrentFSxLustreDeletionPolicy

The current deletion policy for the Amazon FSx for Lustre file system in the shared environment.

Type: String

Valid Values: `DeleteIfNotUsed | Keep`

Required: No

DesiredFSxLustreConfig

The desired Amazon FSx for Lustre file system configuration in the shared environment.

Type: FSxLustreConfig object

Required: No

DesiredFSxLustreDeletionPolicy

The desired deletion policy for the Amazon FSx for Lustre file system in the shared environment.

Type: String

Valid Values: `DeleteIfNotUsed | Keep`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterSlurmConfig

Service: Amazon SageMaker Service

The Slurm configuration for an instance group in a SageMaker HyperPod cluster.

Contents

NodeType

The type of Slurm node for the instance group. Valid values are `Controller`, `Worker`, and `Login`.

Type: String

Valid Values: `Controller | Login | Compute`

Required: Yes

PartitionNames

The list of Slurm partition names that the instance group belongs to.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterSlurmConfigDetails

Service: Amazon SageMaker Service

The Slurm configuration details for an instance group in a SageMaker HyperPod cluster.

Contents

NodeType

The type of Slurm node for the instance group. Valid values are `Controller`, `Worker`, and `Login`.

Type: String

Valid Values: `Controller | Login | Compute`

Required: Yes

PartitionNames

The list of Slurm partition names that the instance group belongs to.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterSpotOptions

Service: Amazon SageMaker Service

Configuration options specific to Spot instances.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterSummary

Service: Amazon SageMaker Service

Lists a summary of the properties of a SageMaker HyperPod cluster.

Contents

ClusterArn

The Amazon Resource Name (ARN) of the SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Required: Yes

ClusterName

The name of the SageMaker HyperPod cluster.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

ClusterStatus

The status of the SageMaker HyperPod cluster.

Type: String

Valid Values: `Creating | Deleting | Failed | InService | RollingBack | SystemUpdating | Updating`

Required: Yes

CreationTime

The time when the SageMaker HyperPod cluster is created.

Type: Timestamp

Required: Yes

TrainingPlanArns

A list of Amazon Resource Names (ARNs) of the training plans associated with this cluster.

For more information about how to reserve GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Type: Array of strings

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterTieredStorageConfig

Service: Amazon SageMaker Service

Defines the configuration for managed tier checkpointing in a HyperPod cluster. Managed tier checkpointing uses multiple storage tiers, including cluster CPU memory, to provide faster checkpoint operations and improved fault tolerance for large-scale model training. The system automatically saves checkpoints at high frequency to memory and periodically persists them to durable storage, like Amazon S3.

Contents

Mode

Specifies whether managed tier checkpointing is enabled or disabled for the HyperPod cluster. When set to `Enable`, the system installs a memory management daemon that provides disaggregated memory as a service for checkpoint storage. When set to `Disable`, the feature is turned off and the memory management daemon is removed from the cluster.

Type: String

Valid Values: `Enable | Disable`

Required: Yes

InstanceMemoryAllocationPercentage

The percentage (int) of cluster memory to allocate for checkpointing.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

CodeEditorAppImageConfig

Service: Amazon SageMaker Service

The configuration for the file system and kernels in a SageMaker image running as a Code Editor app. The `FileSystemConfig` object is not supported.

Contents

ContainerConfig

The configuration used to run the application image container.

Type: ContainerConfig object

Required: No

FileSystemConfig

The Amazon Elastic File System storage configuration for a SageMaker AI image.

Type: FileSystemConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CodeEditorAppSettings

Service: Amazon SageMaker Service

The Code Editor application settings.

For more information about Code Editor, see Get started with Code Editor in Amazon SageMaker.

Contents

AppLifecycleManagement

Settings that are used to configure and manage the lifecycle of CodeEditor applications.

Type: AppLifecycleManagement object

Required: No

BuiltInLifecycleConfigArn

The lifecycle configuration that runs before the default lifecycle configuration. It can override changes made in the default lifecycle configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:studio-lifecycle-config/.*|None)`

Required: No

CustomImages

A list of custom SageMaker images that are configured to run as a Code Editor app.

Type: Array of CustomImage objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

DefaultResourceSpec

Specifies the ARN's of a SageMaker AI image and SageMaker AI image version, and the instance type that the version runs on.

> **Note**
>
> When both `SageMakerImageVersionArn` and `SageMakerImageArn` are passed, `SageMakerImageVersionArn` is used. Any updates to `SageMakerImageArn` will not take effect if `SageMakerImageVersionArn` already exists in the `ResourceSpec` because `SageMakerImageVersionArn` always takes precedence. To clear the value set for `SageMakerImageVersionArn`, pass `None` as the value.

Type: ResourceSpec object

Required: No

LifecycleConfigArns

The Amazon Resource Name (ARN) of the Code Editor application lifecycle configuration.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:studio-lifecycle-config/.*|None)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CodeRepository

Service: Amazon SageMaker Service

A Git repository that SageMaker AI automatically displays to users for cloning in the JupyterServer application.

Contents

RepositoryUrl

The URL of the Git repository.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `https://([.\-_a-zA-Z0-9]+/?){3,1016}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CodeRepositorySummary

Service: Amazon SageMaker Service

Specifies summary information about a Git repository.

Contents

CodeRepositoryArn

The Amazon Resource Name (ARN) of the Git repository.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:code-repository/[\S]{1,2048}`

Required: Yes

CodeRepositoryName

The name of the Git repository.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

CreationTime

The date and time that the Git repository was created.

Type: Timestamp

Required: Yes

LastModifiedTime

The date and time that the Git repository was last modified.

Type: Timestamp

Required: Yes

GitConfig

Configuration details for the Git repository, including the URL where it is located and the ARN of the AWS Secrets Manager secret that contains the credentials used to access the repository.

Type: GitConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CognitoConfig

Service: Amazon SageMaker Service

Use this parameter to configure your Amazon Cognito workforce. A single Cognito workforce is created using and corresponds to a single Amazon Cognito user pool.

Contents

ClientId

The client ID for your Amazon Cognito user pool.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[ -~]+`

Required: Yes

UserPool

A user pool is a user directory in Amazon Cognito. With a user pool, your users can sign in to your web or mobile app through Amazon Cognito. Your users can also sign in through social identity providers like Google, Facebook, Amazon, or Apple, and through SAML identity providers.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 55.

Pattern: `[\w-]+_[0-9a-zA-Z]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

CognitoMemberDefinition

Service: Amazon SageMaker Service

Identifies a Amazon Cognito user group. A user group can be used in on or more work teams.

Contents

ClientId

An identifier for an application client. You must create the app client ID using Amazon Cognito.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[ -~]+`

Required: Yes

UserGroup

An identifier for a user group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\p{L}\p{M}\p{S}\p{N}\p{P}]+`

Required: Yes

UserPool

An identifier for a user pool. The user pool must be in the same region as the service that you are calling.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 55.

Pattern: `[\w-]+_[0-9a-zA-Z]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CollectionConfig

Service: Amazon SageMaker Service

Configuration for your collection.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

VectorConfig

Configuration for your vector collection type.

- `Dimension`: The number of elements in your vector.

Type: VectorConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CollectionConfiguration

Service: Amazon SageMaker Service

Configuration information for the Amazon SageMaker Debugger output tensor collections.

Contents

CollectionName

The name of the tensor collection. The name must be unique relative to other rule configuration names.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*`

Required: No

CollectionParameters

Parameter values for the tensor collection. The allowed parameters are `"name"`, `"include_regex"`, `"reduction_config"`, `"save_config"`, `"tensor_names"`, and `"save_histogram"`.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 20 items.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CompilationJobSummary

Service: Amazon SageMaker Service

A summary of a model compilation job.

Contents

CompilationJobArn

The Amazon Resource Name (ARN) of the model compilation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:compilation-job/.*`

Required: Yes

CompilationJobName

The name of the model compilation job that you want a summary for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

CompilationJobStatus

The status of the model compilation job.

Type: String

Valid Values: `INPROGRESS | COMPLETED | FAILED | STARTING | STOPPING | STOPPED`

Required: Yes

CreationTime

The time when the model compilation job was created.

Type: Timestamp

Required: Yes

CompilationEndTime

The time when the model compilation job completed.

Type: Timestamp

Required: No

CompilationStartTime

The time when the model compilation job started.

Type: Timestamp

Required: No

CompilationTargetDevice

The type of device that the model will run on after the compilation job has completed.

Type: String

Valid Values: `lambda | ml_m4 | ml_m5 | ml_m6g | ml_c4 | ml_c5 | ml_c6g | ml_p2 | ml_p3 | ml_g4dn | ml_inf1 | ml_inf2 | ml_trn1 | ml_eia2 | jetson_tx1 | jetson_tx2 | jetson_nano | jetson_xavier | rasp3b | rasp4b | imx8qm | deeplens | rk3399 | rk3288 | aisage | sbe_c | qcs605 | qcs603 | sitara_am57x | amba_cv2 | amba_cv22 | amba_cv25 | x86_win32 | x86_win64 | coreml | jacinto_tda4vm | imx8mplus`

Required: No

CompilationTargetPlatformAccelerator

The type of accelerator that the model will run on after the compilation job has completed.

Type: String

Valid Values: `INTEL_GRAPHICS | MALI | NVIDIA | NNA`

Required: No

CompilationTargetPlatformArch

The type of architecture that the model will run on after the compilation job has completed.

Type: String

Valid Values: `X86_64 | X86 | ARM64 | ARM_EABI | ARM_EABIHF`

Required: No

CompilationTargetPlatformOs

The type of OS that the model will run on after the compilation job has completed.

Type: String

Valid Values: `ANDROID | LINUX`

Required: No

LastModifiedTime

The time when the model compilation job was last modified.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ComputeQuotaConfig

Service: Amazon SageMaker Service

Configuration of the compute allocation definition for an entity. This includes the resource sharing option and the setting to preempt low priority tasks.

Contents

ComputeQuotaResources

Allocate compute resources by instance types.

Type: Array of ComputeQuotaResourceConfig objects

Array Members: Minimum number of 0 items. Maximum number of 15 items.

Required: No

PreemptTeamTasks

Allows workloads from within an entity to preempt same-team workloads. When set to `LowerPriority`, the entity's lower priority tasks are preempted by their own higher priority tasks.

Default is `LowerPriority`.

Type: String

Valid Values: `Never | LowerPriority`

Required: No

ResourceSharingConfig

Resource sharing configuration. This defines how an entity can lend and borrow idle compute with other entities within the cluster.

Type: ResourceSharingConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ComputeQuotaResourceConfig

Service: Amazon SageMaker Service

Configuration of the resources used for the compute allocation definition.

Contents

InstanceType

The instance type of the instance group for the cluster.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge | ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge`

| ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge

Required: Yes

AcceleratorPartition

The accelerator partition configuration for fractional GPU allocation.

Type: AcceleratorPartitionConfig object

Required: No

Accelerators

The number of accelerators to allocate. If you don't specify a value for vCPU and MemoryInGiB, SageMaker AI automatically allocates ratio-based values for those parameters based on the number of accelerators you provide. For example, if you allocate 16 out of 32 total accelerators, SageMaker AI uses the ratio of 0.5 and allocates values to vCPU and MemoryInGiB.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

Count

The number of instances to add to the instance group of a SageMaker HyperPod cluster.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

MemoryInGiB

The amount of memory in GiB to allocate. If you specify a value only for this parameter, SageMaker AI automatically allocates a ratio-based value for vCPU based on this memory that

you provide. For example, if you allocate 200 out of 400 total memory in GiB, SageMaker AI uses the ratio of 0.5 and allocates values to vCPU. Accelerators are set to 0.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

VCpu

The number of vCPU to allocate. If you specify a value only for vCPU, SageMaker AI automatically allocates ratio-based values for MemoryInGiB based on this vCPU parameter. For example, if you allocate 20 out of 40 total vCPU, SageMaker AI uses the ratio of 0.5 and allocates values to MemoryInGiB. Accelerators are set to 0.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ComputeQuotaSummary

Service: Amazon SageMaker Service

Summary of the compute allocation definition.

Contents

ComputeQuotaArn

ARN of the compute allocation definition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:compute-quota/[a-z0-9]{12}`

Required: Yes

ComputeQuotaId

ID of the compute allocation definition.

Type: String

Pattern: `[a-z0-9]{12}`

Required: Yes

ComputeQuotaTarget

The target entity to allocate compute resources to.

Type: ComputeQuotaTarget object

Required: Yes

CreationTime

Creation time of the compute allocation definition.

Type: Timestamp

Required: Yes

Name

Name of the compute allocation definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Status

Status of the compute allocation definition.

Type: String

Valid Values: `Creating | CreateFailed | CreateRollbackFailed | Created | Updating | UpdateFailed | UpdateRollbackFailed | Updated | Deleting | DeleteFailed | DeleteRollbackFailed | Deleted`

Required: Yes

ActivationState

The state of the compute allocation being described. Use to enable or disable compute allocation.

Default is `Enabled`.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

ClusterArn

ARN of the cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:cluster/[a-z0-9]{12}`

Required: No

ComputeQuotaConfig

Configuration of the compute allocation definition. This includes the resource sharing option, and the setting to preempt low priority tasks.

Type: ComputeQuotaConfig object

Required: No

ComputeQuotaVersion

Version of the compute allocation definition.

Type: Integer

Required: No

LastModifiedTime

Last modified time of the compute allocation definition.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ComputeQuotaTarget

Service: Amazon SageMaker Service

The target entity to allocate compute resources to.

Contents

TeamName

Name of the team to allocate compute resources to.

Type: String

Pattern: `[a-z0-9]([-a-z0-9]*[a-z0-9]){0,39}?`

Required: Yes

FairShareWeight

Assigned entity fair-share weight. Idle compute will be shared across entities based on these assigned weights. This weight is only used when `FairShare` is enabled.

A weight of 0 is the lowest priority and 100 is the highest. Weight 0 is the default.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConditionStepMetadata

Service: Amazon SageMaker Service

Metadata for a Condition step.

Contents

Outcome

The outcome of the Condition step evaluation.

Type: String

Valid Values: `True | False`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContainerConfig

Service: Amazon SageMaker Service

The configuration used to run the application image container.

Contents

ContainerArguments

The arguments for the container when you're running the application.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `(?!\s*$).+`

Required: No

ContainerEntrypoint

The entrypoint used to run the application in the container.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(?!\s*$).+`

Required: No

ContainerEnvironmentVariables

The environment variables to set in the container

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 25 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `(?!\s*$).+`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContainerDefinition

Service: Amazon SageMaker Service

Describes the container, as part of model definition.

Contents

AdditionalModelDataSources

Data sources that are available to your model in addition to the one that you specify for `ModelDataSource` when you use the `CreateModel` action.

Type: Array of AdditionalModelDataSource objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

ContainerHostname

This parameter is ignored for models that contain only a `PrimaryContainer`.

When a `ContainerDefinition` is part of an inference pipeline, the value of the parameter uniquely identifies the container for the purposes of logging and metrics. For information, see Use Logs and Metrics to Monitor an Inference Pipeline. If you don't specify a value for this parameter for a `ContainerDefinition` that is part of an inference pipeline, a unique name is automatically assigned based on the position of the `ContainerDefinition` in the pipeline. If you specify a value for the `ContainerHostName` for any `ContainerDefinition` that is part of an inference pipeline, you must specify a value for the `ContainerHostName` parameter of every `ContainerDefinition` in that pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ContainerMetricsConfig

The configuration for container metrics scraping. Specifies the metrics endpoint path and publishing frequency. If not specified when `EnableDetailedObservability` is `True`, the

default path `/metrics` on port `8080` is used. For first-party and Deep Learning Containers (DLC), the endpoint path is determined automatically and this configuration is optional.

Type: ContainerMetricsConfig object

Required: No

Environment

The environment variables to set in the Docker container. Don't include any sensitive data in your environment variables.

The maximum length of each key and value in the `Environment` map is 1024 bytes. The maximum length of all keys and values in the map, combined, is 32 KB. If you pass multiple containers to a `CreateModel` request, then the maximum length of all of their maps, combined, is also 32 KB.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 1024.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 1024.

Value Pattern: `[\S\s]*`

Required: No

Image

The path where inference code is stored. This can be either in Amazon EC2 Container Registry or in a Docker registry that is accessible from the same VPC that you configure for your endpoint. If you are using your own custom algorithm instead of an algorithm provided by SageMaker, the inference code must meet SageMaker requirements. SageMaker supports both `registry/repository[:tag]` and `registry/repository[@digest]` image path formats. For more information, see Using Your Own Algorithms with Amazon SageMaker.

> **Note**
>
> The model artifacts in an Amazon S3 bucket and the Docker image for inference container in Amazon EC2 Container Registry must be in the same region as the model or endpoint you are creating.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[\S]+`

Required: No

ImageConfig

Specifies whether the model container is in Amazon ECR or a private Docker registry accessible from your Amazon Virtual Private Cloud (VPC). For information about storing containers in a private Docker registry, see Use a Private Docker Registry for Real-Time Inference Containers.

> **Note**
>
> The model artifacts in an Amazon S3 bucket and the Docker image for inference container in Amazon EC2 Container Registry must be in the same region as the model or endpoint you are creating.

Type: ImageConfig object

Required: No

InferenceSpecificationName

The inference specification name in the model package version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Mode

Whether the container hosts a single model or multiple models.

Type: String

Valid Values: `SingleModel | MultiModel`

Required: No

ModelDataSource

Specifies the location of ML model data to deploy.

> **Note**
>
> Currently you cannot use `ModelDataSource` in conjunction with SageMaker batch transform, SageMaker serverless endpoints, SageMaker multi-model endpoints, and SageMaker Marketplace.

Type: ModelDataSource object

Required: No

ModelDataUrl

The S3 path where the model artifacts, which result from model training, are stored. This path must point to a single gzip compressed tar archive (.tar.gz suffix). The S3 path is required for SageMaker built-in algorithms, but not if you use your own algorithms. For more information on built-in algorithms, see Common Parameters.

> **Note**
>
> The model artifacts must be in an S3 bucket that is in the same region as the model or endpoint you are creating.

If you provide a value for this parameter, SageMaker uses AWS Security Token Service to download model artifacts from the S3 path you provide. AWS STS is activated in your AWS account by default. If you previously deactivated AWS STS for a region, you need to reactivate

AWS STS for that region. For more information, see Activating and Deactivating AWS STS in an AWS Region in the *AWS Identity and Access Management User Guide*.

> **Important**
>
> If you use a built-in algorithm to create a model, SageMaker requires that you provide a S3 path to the model artifacts in `ModelDataUrl`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

ModelPackageName

The name or Amazon Resource Name (ARN) of the model package to use to create the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 176.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*\/)?([a-zA-Z0-9]([a-zA-Z0-9-]){0,62})(?<!-)(\/[0-9]{1,9})?`

Required: No

MultiModelConfig

Specifies additional configuration for multi-model endpoints.

Type: MultiModelConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContainerMetricsConfig

Service: Amazon SageMaker Service

The configuration for container-level metrics scraping. Use this configuration to specify a custom metrics endpoint path and publishing frequency for container metrics. When `EnableDetailedObservability` is set to `True` in `MetricsConfig`, metrics are scraped from the container's Prometheus endpoint. If this configuration is not provided, the default path `/metrics` on port `8080` is used with a default publishing frequency of `60` seconds. For first-party and Deep Learning Containers (DLC), the endpoint path is determined automatically and this configuration is optional.

Contents

MetricsEndpoints

A list of metrics endpoints to scrape from the container. Each endpoint specifies the path where the container exposes Prometheus-formatted metrics and the frequency at which to publish them. You can specify a maximum of 1 endpoint.

Type: Array of MetricsEndpoint objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContextSource

Service: Amazon SageMaker Service

A structure describing the source of a context.

Contents

SourceUri

The URI of the source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*`

Required: Yes

SourceId

The ID of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

SourceType

The type of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContextSummary

Service: Amazon SageMaker Service

Lists a summary of the properties of a context. A context provides a logical grouping of other entities.

Contents

ContextArn

The Amazon Resource Name (ARN) of the context.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:context/.*`

Required: No

ContextName

The name of the context.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,119}`

Required: No

ContextType

The type of the context.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

CreationTime

When the context was created.

Type: Timestamp

Required: No

LastModifiedTime

When the context was last modified.

Type: Timestamp

Required: No

Source

The source of the context.

Type: ContextSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContinuousParameterRange

Service: Amazon SageMaker Service

A list of continuous hyperparameters to tune.

Contents

MaxValue

The maximum value for the hyperparameter. The tuning job uses floating-point values between `MinValue` value and this value for tuning.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

MinValue

The minimum value for the hyperparameter. The tuning job uses floating-point values between this value and `MaxValue`for tuning.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

Name

The name of the continuous hyperparameter to tune.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

ScalingType

The scale that hyperparameter tuning uses to search the hyperparameter range. For information about choosing a hyperparameter scale, see Hyperparameter Scaling. One of the following values:

Auto

SageMaker hyperparameter tuning chooses the best scale for the hyperparameter.

Linear

Hyperparameter tuning searches the values in the hyperparameter range by using a linear scale.

Logarithmic

Hyperparameter tuning searches the values in the hyperparameter range by using a logarithmic scale.

Logarithmic scaling works only for ranges that have only values greater than 0.

ReverseLogarithmic

Hyperparameter tuning searches the values in the hyperparameter range by using a reverse logarithmic scale.

Reverse logarithmic scaling works only for ranges that are entirely within the range 0<=x<1.0.

Type: String

Valid Values: `Auto | Linear | Logarithmic | ReverseLogarithmic`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ContinuousParameterRangeSpecification

Service: Amazon SageMaker Service

Defines the possible values for a continuous hyperparameter.

Contents

MaxValue

The maximum floating-point value allowed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

MinValue

The minimum floating-point value allowed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConvergenceDetected

Service: Amazon SageMaker Service

A flag to indicating that automatic model tuning (AMT) has detected model convergence, defined as a lack of significant improvement (1% or less) against an objective metric.

Contents

CompleteOnConvergence

A flag to stop a tuning job once AMT has detected that the job has converged.

Type: String

Valid Values: `Disabled | Enabled`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CreateTemplateProvider

Service: Amazon SageMaker Service

Contains configuration details for a template provider. Only one type of template provider can be specified.

Contents

CfnTemplateProvider

The CloudFormation template provider configuration for creating infrastructure resources.

Type: CfnCreateTemplateProvider object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomFileSystem

Service: Amazon SageMaker Service

A file system, created by you, that you assign to a user profile or space for an Amazon SageMaker AI Domain. Permitted users can access this file system in Amazon SageMaker AI Studio.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

EFSFileSystem

A custom file system in Amazon EFS.

Type: EFSFileSystem object

Required: No

FSxLustreFileSystem

A custom file system in Amazon FSx for Lustre.

Type: FSxLustreFileSystem object

Required: No

S3FileSystem

A custom file system in Amazon S3. This is only supported in Amazon SageMaker Unified Studio.

Type: S3FileSystem object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomFileSystemConfig

Service: Amazon SageMaker Service

The settings for assigning a custom file system to a user profile or space for an Amazon SageMaker AI Domain. Permitted users can access this file system in Amazon SageMaker AI Studio.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

EFSFileSystemConfig

The settings for a custom Amazon EFS file system.

Type: EFSFileSystemConfig object

Required: No

FSxLustreFileSystemConfig

The settings for a custom Amazon FSx for Lustre file system.

Type: FSxLustreFileSystemConfig object

Required: No

S3FileSystemConfig

Configuration settings for a custom Amazon S3 file system.

Type: S3FileSystemConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomImage

Service: Amazon SageMaker Service

A custom SageMaker AI image. For more information, see Bring your own SageMaker AI image.

Contents

AppImageConfigName

The name of the AppImageConfig.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ImageName

The name of the CustomImage. Must be unique to your account.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

ImageVersionNumber

The version number of the CustomImage.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomizedMetricSpecification

Service: Amazon SageMaker Service

A customized metric.

Contents

MetricName

The name of the customized metric.

Type: String

Required: No

Namespace

The namespace of the customized metric.

Type: String

Required: No

Statistic

The statistic of the customized metric.

Type: String

Valid Values: `Average | Minimum | Maximum | SampleCount | Sum`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomPosixUserConfig

Service: Amazon SageMaker Service

Details about the POSIX identity that is used for file system operations.

Contents

Gid

The POSIX group ID.

Type: Long

Valid Range: Minimum value of 1001. Maximum value of 4000000.

Required: Yes

Uid

The POSIX user ID.

Type: Long

Valid Range: Minimum value of 10000. Maximum value of 4000000.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataCaptureConfig

Service: Amazon SageMaker Service

Configuration to control how SageMaker AI captures inference data.

Contents

CaptureOptions

Specifies data Model Monitor will capture. You can configure whether to collect only input, only output, or both

Type: Array of CaptureOption objects

Array Members: Minimum number of 1 item. Maximum number of 32 items.

Required: Yes

DestinationS3Uri

The Amazon S3 location used to capture the data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `(https|s3)://([^/])/?(.*)`

Required: Yes

InitialSamplingPercentage

The percentage of requests SageMaker AI will capture. A lower value is recommended for Endpoints with high traffic.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 100.

Required: Yes

CaptureContentTypeHeader

Configuration specifying how to treat different headers. If no headers are specified SageMaker AI will by default base64 encode when capturing the data.

Type: CaptureContentTypeHeader object

Required: No

EnableCapture

Whether data capture should be enabled or disabled (defaults to enabled).

Type: Boolean

Required: No

KmsKeyId

The Amazon Resource Name (ARN) of an AWS Key Management Service key that SageMaker AI uses to encrypt the captured data at rest using Amazon S3 server-side encryption.

The KmsKeyId can be any of the following formats:

- Key ID: `1234abcd-12ab-34cd-56ef-1234567890ab`
- Key ARN: `arn:aws:kms:us-west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab`
- Alias name: `alias/ExampleAlias`
- Alias name ARN: `arn:aws:kms:us-west-2:111122223333:alias/ExampleAlias`

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataCaptureConfigSummary

Service: Amazon SageMaker Service

The currently active data capture configuration used by your Endpoint.

Contents

CaptureStatus

Whether data capture is currently functional.

Type: String

Valid Values: `Started | Stopped`

Required: Yes

CurrentSamplingPercentage

The percentage of requests being captured by your Endpoint.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 100.

Required: Yes

DestinationS3Uri

The Amazon S3 location being used to capture the data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `(https|s3)://([^/])/?(.*)`

Required: Yes

EnableCapture

Whether data capture is enabled or disabled.

Type: Boolean

Required: Yes

KmsKeyId

The KMS key being used to encrypt the data in Amazon S3.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataCatalogConfig

Service: Amazon SageMaker Service

The meta data of the Glue table which serves as data catalog for the `OfflineStore`.

Contents

Catalog

The name of the Glue table catalog.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\t]*`

Required: Yes

Database

The name of the Glue table database.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\t]*`

Required: Yes

TableName

The name of the Glue table.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\t]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataProcessing

Service: Amazon SageMaker Service

The data structure used to specify the data to be used for inference in a batch transform job and to associate the data that is relevant to the prediction results in the output. The input filter provided allows you to exclude input data that is not needed for inference in a batch transform job. The output filter provided allows you to include input data relevant to interpreting the predictions in the output from the job. For more information, see Associate Prediction Results with their Corresponding Input Records.

Contents

InputFilter

A JSONPath expression used to select a portion of the input data to pass to the algorithm. Use the `InputFilter` parameter to exclude fields, such as an ID column, from the input. If you want SageMaker to pass the entire input dataset to the algorithm, accept the default value $.

Examples: `"$"`, `"$[1:]"`, `"$.features"`

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Required: No

JoinSource

Specifies the source of the data to join with the transformed data. The valid values are `None` and `Input`. The default value is `None`, which specifies not to join the input with the transformed data. If you want the batch transform job to join the original input data with the transformed data, set `JoinSource` to `Input`. You can specify `OutputFilter` as an additional filter to select a portion of the joined dataset and store it in the output file.

For JSON or JSONLines objects, such as a JSON array, SageMaker adds the transformed data to the input JSON object in an attribute called `SageMakerOutput`. The joined result for JSON must be a key-value pair object. If the input is not a key-value pair object, SageMaker creates a new JSON file. In the new JSON file, and the input data is stored under the `SageMakerInput` key and the results are stored in `SageMakerOutput`.

For CSV data, SageMaker takes each row as a JSON array and joins the transformed data with the input by appending each transformed row to the end of the input. The joined data has the original input data followed by the transformed data and the output is a CSV file.

For information on how joining in applied, see Workflow for Associating Inferences with Input Records.

Type: String

Valid Values: `Input | None`

Required: No

OutputFilter

A JSONPath expression used to select a portion of the joined dataset to save in the output file for a batch transform job. If you want SageMaker to store the entire input dataset in the output file, leave the default value, $. If you specify indexes that aren't within the dimension size of the joined dataset, you get an error.

Examples: `"$"`, `"$[0,5:]"`, `"$['id','SageMakerOutput']"`

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataQualityAppSpecification

Service: Amazon SageMaker Service

Information about the container that a data quality monitoring job runs.

Contents

ImageUri

The container image that the data quality monitoring job runs.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

ContainerArguments

The arguments to send to the container that the monitoring job runs.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

ContainerEntrypoint

The entrypoint for a container used to run a monitoring job.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

Environment

Sets the environment variables in the container that the monitoring job runs.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[\S\s]*`

Required: No

PostAnalyticsProcessorSourceUri

An Amazon S3 URI to a script that is called after analysis has been performed. Applicable only for the built-in (first party) containers.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

RecordPreprocessorSourceUri

An Amazon S3 URI to a script that is called per row prior to running analysis. It can base64 decode the payload and convert it into a flattened JSON so that the built-in container can use the converted data. Applicable only for the built-in (first party) containers.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataQualityBaselineConfig

Service: Amazon SageMaker Service

Configuration for monitoring constraints and monitoring statistics. These baseline resources are compared against the results of the current job from the series of jobs scheduled to collect data periodically.

Contents

BaseliningJobName

The name of the job that performs baselining for the data quality monitoring job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ConstraintsResource

The constraints resource for a monitoring job.

Type: MonitoringConstraintsResource object

Required: No

StatisticsResource

The statistics resource for a monitoring job.

Type: MonitoringStatisticsResource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataQualityJobInput

Service: Amazon SageMaker Service

The input for the data quality monitoring job. Currently endpoints are supported for input.

Contents

BatchTransformInput

Input object for the batch transform job.

Type: BatchTransformInput object

Required: No

EndpointInput

Input object for the endpoint

Type: EndpointInput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DatasetDefinition

Service: Amazon SageMaker Service

Configuration for Dataset Definition inputs. The Dataset Definition input must specify exactly one of either `AthenaDatasetDefinition` or `RedshiftDatasetDefinition` types.

Contents

AthenaDatasetDefinition

Configuration for Athena Dataset Definition input.

Type: AthenaDatasetDefinition object

Required: No

DataDistributionType

Whether the generated dataset is `FullyReplicated` or `ShardedByS3Key` (default).

Type: String

Valid Values: `FullyReplicated | ShardedByS3Key`

Required: No

InputMode

Whether to use `File` or `Pipe` input mode. In `File` (default) mode, Amazon SageMaker copies the data from the input source onto the local Amazon Elastic Block Store (Amazon EBS) volumes before starting your training algorithm. This is the most commonly used input mode. In `Pipe` mode, Amazon SageMaker streams input data from the source directly to your algorithm without using the EBS volume.

Type: String

Valid Values: `Pipe | File`

Required: No

LocalPath

The local path where you want Amazon SageMaker to download the Dataset Definition inputs to run a processing job. `LocalPath` is an absolute path to the input data. This is a required parameter when `AppManaged` is `False` (default).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

RedshiftDatasetDefinition

Configuration for Redshift Dataset Definition input.

Type: RedshiftDatasetDefinition object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DatasetSource

Service: Amazon SageMaker Service

Specifies a dataset source for a channel.

Contents

DatasetArn

The Amazon Resource Name (ARN) of the dataset resource.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub-content\/)[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}\/DataSet\/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}(\/\d{1,4}.\d{1,4}.\d{1,4})?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataSource

Service: Amazon SageMaker Service

Describes the location of the channel data.

Contents

DatasetSource

The dataset resource that's associated with a channel.

Type: DatasetSource object

Required: No

FileSystemDataSource

The file system that is associated with a channel.

Type: FileSystemDataSource object

Required: No

S3DataSource

The S3 location of the data source that is associated with a channel.

Type: S3DataSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DebugHookConfig

Service: Amazon SageMaker Service

Configuration information for the Amazon SageMaker Debugger hook parameters, metric and tensor collections, and storage paths. To learn more about how to configure the `DebugHookConfig` parameter, see Use the SageMaker and Debugger Configuration API Operations to Create, Update, and Debug Your Training Job.

Contents

S3OutputPath

Path to Amazon S3 storage location for metrics and tensors.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

CollectionConfigurations

Configuration information for Amazon SageMaker Debugger tensor collections. To learn more about how to configure the `CollectionConfiguration` parameter, see Use the SageMaker and Debugger Configuration API Operations to Create, Update, and Debug Your Training Job.

Type: Array of CollectionConfiguration objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

HookParameters

Configuration information for the Amazon SageMaker Debugger hook parameters.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 20 items.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `.*`

Required: No

LocalPath

Path to local storage location for metrics and tensors. Defaults to `/opt/ml/output/tensors/`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4096.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DebugRuleConfiguration

Service: Amazon SageMaker Service

Configuration information for SageMaker Debugger rules for debugging. To learn more about how to configure the `DebugRuleConfiguration` parameter, see Use the SageMaker and Debugger Configuration API Operations to Create, Update, and Debug Your Training Job.

Contents

RuleConfigurationName

The name of the rule configuration. It must be unique relative to other rule configuration names.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*`

Required: Yes

RuleEvaluatorImage

The Amazon Elastic Container (ECR) Image for the managed rule evaluation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

InstanceType

The instance type to deploy a custom rule for debugging a training job.

Type: String

Valid Values: `ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge`

```
| ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge
| ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.xlarge
| ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge
| ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge |
ml.m5.12xlarge | ml.m5.24xlarge | ml.r5.large | ml.r5.xlarge |
ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge |
ml.r5.16xlarge | ml.r5.24xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge |
ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge
| ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge |
ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge
| ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge |
ml.r5d.8xlarge | ml.r5d.12xlarge | ml.r5d.16xlarge | ml.r5d.24xlarge
| ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge |
ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge |
ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge |
ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge
| ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge |
ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge
| ml.m6i.32xlarge | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge |
ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge
| ml.c6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge |
ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge
| ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge
| ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge
| ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large
| ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge |
ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge
| ml.p5.4xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge |
ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge
```

Required: No

LocalPath

Path to local storage location for output of rules. Defaults to `/opt/ml/processing/output/rule/`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4096.

Pattern: `.*`

Required: No

RuleParameters

Runtime configuration for rule container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `.*`

Required: No

S3OutputPath

Path to Amazon S3 storage location for rules.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

VolumeSizeInGB

The size, in GB, of the ML storage volume attached to the processing instance.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DebugRuleEvaluationStatus

Service: Amazon SageMaker Service

Information about the status of the rule evaluation.

Contents

LastModifiedTime

Timestamp when the rule evaluation status was last modified.

Type: Timestamp

Required: No

RuleConfigurationName

The name of the rule configuration.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*`

Required: No

RuleEvaluationJobArn

The Amazon Resource Name (ARN) of the rule evaluation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:processing-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

RuleEvaluationStatus

Status of the rule evaluation.

Type: String

Valid Values: `InProgress | NoIssuesFound | IssuesFound | Error | Stopping | Stopped`

Required: No

StatusDetails

Details from the rule evaluation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DefaultEbsStorageSettings

Service: Amazon SageMaker Service

A collection of default EBS storage settings that apply to spaces created within a domain or user profile.

Contents

DefaultEbsVolumeSizeInGb

The default size of the EBS storage volume for a space.

Type: Integer

Valid Range: Minimum value of 5. Maximum value of 16384.

Required: Yes

MaximumEbsVolumeSizeInGb

The maximum size of the EBS storage volume for a space.

Type: Integer

Valid Range: Minimum value of 5. Maximum value of 16384.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DefaultSpaceSettings

Service: Amazon SageMaker Service

The default settings for shared spaces that users create in the domain.

SageMaker applies these settings only to shared spaces. It doesn't apply them to private spaces.

Contents

CustomFileSystemConfigs

The settings for assigning a custom file system to a domain. Permitted users can access this file system in Amazon SageMaker AI Studio.

Type: Array of CustomFileSystemConfig objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

CustomPosixUserConfig

Details about the POSIX identity that is used for file system operations.

Type: CustomPosixUserConfig object

Required: No

ExecutionRole

The ARN of the execution role for the space.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

JupyterLabAppSettings

The settings for the JupyterLab application.

Type: JupyterLabAppSettings object

Required: No

JupyterServerAppSettings

The JupyterServer app settings.

Type: JupyterServerAppSettings object

Required: No

KernelGatewayAppSettings

The KernelGateway app settings.

Type: KernelGatewayAppSettings object

Required: No

SecurityGroups

The security group IDs for the Amazon VPC that the space uses for communication.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: No

SpaceStorageSettings

The default storage settings for a space.

Type: DefaultSpaceStorageSettings object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

DefaultSpaceStorageSettings

Service: Amazon SageMaker Service

The default storage settings for a space.

Contents

DefaultEbsStorageSettings

The default EBS storage settings for a space.

Type: DefaultEbsStorageSettings object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeployedImage

Service: Amazon SageMaker Service

Gets the Amazon EC2 Container Registry path of the docker image of the model that is hosted in this ProductionVariant.

If you used the `registry/repository[:tag]` form to specify the image path of the primary container when you created the model hosted in this `ProductionVariant`, the path resolves to a path of the form `registry/repository[@digest]`. A digest is a hash value that identifies a specific version of an image. For information about Amazon ECR paths, see Pulling an Image in the *Amazon ECR User Guide*.

Contents

ResolutionTime

The date and time when the image path for the model resolved to the `ResolvedImage`

Type: Timestamp

Required: No

ResolvedImage

The specific digest path of the image hosted in this `ProductionVariant`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[\S]+`

Required: No

SpecifiedImage

The image path you specified when you created the model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[\S]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeploymentConfig

Service: Amazon SageMaker Service

The deployment configuration for an endpoint, which contains the desired deployment strategy and rollback configurations.

Contents

AutoRollbackConfiguration

Automatic rollback configuration for handling endpoint deployment failures and recovery.

Type: AutoRollbackConfig object

Required: No

BlueGreenUpdatePolicy

Update policy for a blue/green deployment. If this update policy is specified, SageMaker creates a new fleet during the deployment while maintaining the old fleet. SageMaker flips traffic to the new fleet according to the specified traffic routing configuration. Only one update policy should be used in the deployment configuration. If no update policy is specified, SageMaker uses a blue/green deployment strategy with all at once traffic shifting by default.

Type: BlueGreenUpdatePolicy object

Required: No

RollingUpdatePolicy

Specifies a rolling deployment strategy for updating a SageMaker endpoint.

Type: RollingUpdatePolicy object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeploymentConfiguration

Service: Amazon SageMaker Service

The configuration to use when updating the AMI versions.

Contents

AutoRollbackConfiguration

An array that contains the alarms that SageMaker monitors to know whether to roll back the AMI update.

Type: Array of AlarmDetails objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

RollingUpdatePolicy

The policy that SageMaker uses when updating the AMI versions of the cluster.

Type: RollingDeploymentPolicy object

Required: No

WaitIntervalInSeconds

The duration in seconds that SageMaker waits before updating more instances in the cluster.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 3600.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

DeploymentRecommendation

Service: Amazon SageMaker Service

A set of recommended deployment configurations for the model. To get more advanced recommendations, see CreateInferenceRecommendationsJob to create an inference recommendation job.

Contents

RecommendationStatus

Status of the deployment recommendation. The status `NOT_APPLICABLE` means that SageMaker is unable to provide a default recommendation for the model using the information provided. If the deployment status is `IN_PROGRESS`, retry your API call after a few seconds to get a `COMPLETED` deployment recommendation.

Type: String

Valid Values: `IN_PROGRESS | COMPLETED | FAILED | NOT_APPLICABLE`

Required: Yes

RealTimeInferenceRecommendations

A list of RealTimeInferenceRecommendation items.

Type: Array of RealTimeInferenceRecommendation objects

Array Members: Minimum number of 0 items. Maximum number of 3 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeploymentStage

Service: Amazon SageMaker Service

Contains information about a stage in an edge deployment plan.

Contents

DeviceSelectionConfig

Configuration of the devices in the stage.

Type: DeviceSelectionConfig object

Required: Yes

StageName

The name of the stage.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

DeploymentConfig

Configuration of the deployment details.

Type: EdgeDeploymentConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeploymentStageStatusSummary

Service: Amazon SageMaker Service

Contains information summarizing the deployment stage results.

Contents

DeploymentConfig

Configuration of the deployment details.

Type: EdgeDeploymentConfig object

Required: Yes

DeploymentStatus

General status of the current state.

Type: EdgeDeploymentStatus object

Required: Yes

DeviceSelectionConfig

Configuration of the devices in the stage.

Type: DeviceSelectionConfig object

Required: Yes

StageName

The name of the stage.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DerivedInformation

Service: Amazon SageMaker Service

Information that SageMaker Neo automatically derived about the model.

Contents

DerivedDataInputConfig

The data input configuration that SageMaker Neo automatically derived for the model. When SageMaker Neo derives this information, you don't need to specify the data input configuration when you create a compilation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 16384.

Pattern: `[\S\s]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DesiredWeightAndCapacity

Service: Amazon SageMaker Service

Specifies weight and capacity values for a production variant.

Contents

VariantName

The name of the variant to update.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

DesiredInstanceCount

The variant's capacity.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

DesiredWeight

The variant's weight.

Type: Float

Valid Range: Minimum value of 0.

Required: No

ServerlessUpdateConfig

Specifies the serverless update concurrency configuration for an endpoint variant.

Type: ProductionVariantServerlessUpdateConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Device

Service: Amazon SageMaker Service

Information of a particular device.

Contents

DeviceName

The name of the device.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Description

Description of the device.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 40.

Pattern: `[-a-zA-Z0-9_.,;:! ]*`

Required: No

IotThingName

AWS Internet of Things (IoT) object name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `[a-zA-Z0-9:_-]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeviceDeploymentSummary

Service: Amazon SageMaker Service

Contains information summarizing device details and deployment status.

Contents

DeviceArn

The ARN of the device.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:[a-z\-]*:[a-z\-]*:\d{12}:[a-z\-]*/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

DeviceName

The name of the device.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EdgeDeploymentPlanArn

The ARN of the edge deployment plan.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z\-]*:\d{12}:edge-deployment/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

EdgeDeploymentPlanName

The name of the edge deployment plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

StageName

The name of the stage in the edge deployment plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

DeployedStageName

The name of the deployed stage.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

DeploymentStartTime

The time when the deployment on the device started.

Type: Timestamp

Required: No

Description

The description of the device.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 40.

Pattern: `[-a-zA-Z0-9_.,;:! ]*`

Required: No

DeviceDeploymentStatus

The deployment status of the device.

Type: String

Valid Values: `READYTODEPLOY | INPROGRESS | DEPLOYED | FAILED | STOPPING | STOPPED`

Required: No

DeviceDeploymentStatusMessage

The detailed error message for the deployoment status result.

Type: String

Required: No

DeviceFleetName

The name of the fleet to which the device belongs to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeviceFleetSummary

Service: Amazon SageMaker Service

Summary of the device fleet.

Contents

DeviceFleetArn

Amazon Resource Name (ARN) of the device fleet.

Type: String

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:device-fleet/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

DeviceFleetName

Name of the device fleet.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

CreationTime

Timestamp of when the device fleet was created.

Type: Timestamp

Required: No

LastModifiedTime

Timestamp of when the device fleet was last updated.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeviceSelectionConfig

Service: Amazon SageMaker Service

Contains information about the configurations of selected devices.

Contents

DeviceSubsetType

Type of device subsets to deploy to the current stage.

Type: String

Valid Values: `PERCENTAGE | SELECTION | NAMECONTAINS`

Required: Yes

DeviceNameContains

A filter to select devices with names containing this name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

DeviceNames

List of devices chosen to deploy.

Type: Array of strings

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Percentage

Percentage of devices in the fleet to deploy to the current stage.

Type: Integer

Valid Range: Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeviceStats

Service: Amazon SageMaker Service

Status of devices.

Contents

ConnectedDeviceCount

The number of devices connected with a heartbeat.

Type: Long

Required: Yes

RegisteredDeviceCount

The number of registered devices.

Type: Long

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeviceSummary

Service: Amazon SageMaker Service

Summary of the device.

Contents

DeviceArn

Amazon Resource Name (ARN) of the device.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:[a-z\-]*:[a-z\-]*:\d{12}:[a-z\-]*/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

DeviceName

The unique identifier of the device.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AgentVersion

Edge Manager agent version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 30.

Pattern: `[a-zA-Z0-9\ \_\.]+`

Required: No

Description

A description of the device.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 40.

Pattern: `[-a-zA-Z0-9_.,;:! ]*`

Required: No

DeviceFleetName

The name of the fleet the device belongs to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

IotThingName

The AWS Internet of Things (IoT) object thing name associated with the device..

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `[a-zA-Z0-9:_-]+`

Required: No

LatestHeartbeat

The last heartbeat received from the device.

Type: Timestamp

Required: No

Models

Models on the device.

Type: Array of EdgeModelSummary objects

Required: No

RegistrationTime

The timestamp of the last registration or de-reregistration.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DirectDeploySettings

Service: Amazon SageMaker Service

The model deployment settings for the SageMaker Canvas application.

> **Note**
>
> In order to enable model deployment for Canvas, the SageMaker Domain's or user profile's AWS IAM execution role must have the `AmazonSageMakerCanvasDirectDeployAccess` policy attached. You can also turn on model deployment permissions through the SageMaker Domain's or user profile's settings in the SageMaker console.

Contents

Status

Describes whether model deployment permissions are enabled or disabled in the Canvas application.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DockerSettings

Service: Amazon SageMaker Service

A collection of settings that configure the domain's Docker interaction.

Contents

EnableDockerAccess

Indicates whether the domain can access Docker.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

RootlessDocker

Indicates whether to use rootless Docker.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

VpcOnlyTrustedAccounts

The list of AWS accounts that are trusted when the domain is created in VPC-only mode.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Length Constraints: Fixed length of 12.

Pattern: `\d+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DomainDetails

Service: Amazon SageMaker Service

The domain's details.

Contents

CreationTime

The creation time.

Type: Timestamp

Required: No

DomainArn

The domain's Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:domain/.*`

Required: No

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: No

DomainName

The domain name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

LastModifiedTime

The last modified time.

Type: Timestamp

Required: No

Status

The status.

Type: String

Valid Values: `Deleting | Failed | InService | Pending | Updating | Update_Failed | Delete_Failed`

Required: No

Url

The domain's URL.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DomainSettings

Service: Amazon SageMaker Service

A collection of settings that apply to the `SageMaker Domain`. These settings are specified through the `CreateDomain` API call.

Contents

AmazonQSettings

A collection of settings that configure the Amazon Q experience within the domain. The `AuthMode` that you use to create the domain must be `SSO`.

Type: AmazonQSettings object

Required: No

DockerSettings

A collection of settings that configure the domain's Docker interaction.

Type: DockerSettings object

Required: No

ExecutionRoleIdentityConfig

The configuration for attaching a SageMaker AI user profile name to the execution role as a sts:SourceIdentity key.

Type: String

Valid Values: `USER_PROFILE_NAME | DISABLED`

Required: No

IpAddressType

The IP address type for the domain. Specify `ipv4` for IPv4-only connectivity or `dualstack` for both IPv4 and IPv6 connectivity. When you specify `dualstack`, the subnet must support IPv6 CIDR blocks. If not specified, defaults to `ipv4`.

Type: String

Valid Values: `ipv4 | dualstack`

Required: No

RStudioServerProDomainSettings

A collection of settings that configure the `RStudioServerPro` Domain-level app.

Type: RStudioServerProDomainSettings object

Required: No

SecurityGroupIds

The security groups for the Amazon Virtual Private Cloud that the `Domain` uses for communication between Domain-level apps and user apps.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 3 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: No

TrustedIdentityPropagationSettings

The Trusted Identity Propagation (TIP) settings for the SageMaker domain. These settings determine how user identities from IAM Identity Center are propagated through the domain to TIP enabled AWS services.

Type: TrustedIdentityPropagationSettings object

Required: No

UnifiedStudioSettings

The settings that apply to an SageMaker AI domain when you use it in Amazon SageMaker Unified Studio.

Type: UnifiedStudioSettings object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DomainSettingsForUpdate

Service: Amazon SageMaker Service

A collection of `Domain` configuration settings to update.

Contents

AmazonQSettings

A collection of settings that configure the Amazon Q experience within the domain.

Type: AmazonQSettings object

Required: No

DockerSettings

A collection of settings that configure the domain's Docker interaction.

Type: DockerSettings object

Required: No

ExecutionRoleIdentityConfig

The configuration for attaching a SageMaker AI user profile name to the execution role as a sts:SourceIdentity key. This configuration can only be modified if there are no apps in the `InService` or `Pending` state.

Type: String

Valid Values: `USER_PROFILE_NAME | DISABLED`

Required: No

IpAddressType

The IP address type for the domain. Specify `ipv4` for IPv4-only connectivity or `dualstack` for both IPv4 and IPv6 connectivity. When you specify `dualstack`, the subnet must support IPv6 CIDR blocks.

Type: String

Valid Values: `ipv4 | dualstack`

Required: No

RStudioServerProDomainSettingsForUpdate

A collection of `RStudioServerPro` Domain-level app settings to update. A single `RStudioServerPro` application is created for a domain.

Type: RStudioServerProDomainSettingsForUpdate object

Required: No

SecurityGroupIds

The security groups for the Amazon Virtual Private Cloud that the `Domain` uses for communication between Domain-level apps and user apps.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 3 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: No

TrustedIdentityPropagationSettings

The Trusted Identity Propagation (TIP) settings for the SageMaker domain. These settings determine how user identities from IAM Identity Center are propagated through the domain to TIP enabled AWS services.

Type: TrustedIdentityPropagationSettings object

Required: No

UnifiedStudioSettings

The settings that apply to an SageMaker AI domain when you use it in Amazon SageMaker Unified Studio.

Type: UnifiedStudioSettings object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DriftCheckBaselines

Service: Amazon SageMaker Service

Represents the drift check baselines that can be used when the model monitor is set using the model package.

Contents

Bias

Represents the drift check bias baselines that can be used when the model monitor is set using the model package.

Type: DriftCheckBias object

Required: No

Explainability

Represents the drift check explainability baselines that can be used when the model monitor is set using the model package.

Type: DriftCheckExplainability object

Required: No

ModelDataQuality

Represents the drift check model data quality baselines that can be used when the model monitor is set using the model package.

Type: DriftCheckModelDataQuality object

Required: No

ModelQuality

Represents the drift check model quality baselines that can be used when the model monitor is set using the model package.

Type: DriftCheckModelQuality object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DriftCheckBias

Service: Amazon SageMaker Service

Represents the drift check bias baselines that can be used when the model monitor is set using the model package.

Contents

ConfigFile

The bias config file for a model.

Type: FileSource object

Required: No

PostTrainingConstraints

The post-training constraints.

Type: MetricsSource object

Required: No

PreTrainingConstraints

The pre-training constraints.

Type: MetricsSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DriftCheckExplainability

Service: Amazon SageMaker Service

Represents the drift check explainability baselines that can be used when the model monitor is set using the model package.

Contents

ConfigFile

The explainability config file for the model.

Type: FileSource object

Required: No

Constraints

The drift check explainability constraints.

Type: MetricsSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DriftCheckModelDataQuality

Service: Amazon SageMaker Service

Represents the drift check data quality baselines that can be used when the model monitor is set using the model package.

Contents

Constraints

The drift check model data quality constraints.

Type: MetricsSource object

Required: No

Statistics

The drift check model data quality statistics.

Type: MetricsSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DriftCheckModelQuality

Service: Amazon SageMaker Service

Represents the drift check model quality baselines that can be used when the model monitor is set using the model package.

Contents

Constraints

The drift check model quality constraints.

Type: MetricsSource object

Required: No

Statistics

The drift check model quality statistics.

Type: MetricsSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DynamicScalingConfiguration

Service: Amazon SageMaker Service

An object with the recommended values for you to specify when creating an autoscaling policy.

Contents

MaxCapacity

The recommended maximum capacity to specify for your autoscaling policy.

Type: Integer

Required: No

MinCapacity

The recommended minimum capacity to specify for your autoscaling policy.

Type: Integer

Required: No

ScaleInCooldown

The recommended scale in cooldown time for your autoscaling policy.

Type: Integer

Required: No

ScaleOutCooldown

The recommended scale out cooldown time for your autoscaling policy.

Type: Integer

Required: No

ScalingPolicies

An object of the scaling policies for each metric.

Type: Array of ScalingPolicy objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EbsStorageSettings

Service: Amazon SageMaker Service

A collection of EBS storage settings that apply to both private and shared spaces.

Contents

EbsVolumeSizeInGb

The size of an EBS storage volume for a space.

Type: Integer

Valid Range: Minimum value of 5. Maximum value of 16384.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Ec2CapacityReservation

Service: Amazon SageMaker Service

The EC2 capacity reservations that are shared to an ML capacity reservation.

Contents

AvailableInstanceCount

The number of instances that are currently available in the EC2 capacity reservation.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

Ec2CapacityReservationId

The unique identifier for an EC2 capacity reservation that's part of the ML capacity reservation.

Type: String

Required: No

TotalInstanceCount

The number of instances that you allocated to the EC2 capacity reservation.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

UsedByCurrentEndpoint

The number of instances from the EC2 capacity reservation that are being used by the endpoint.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Edge

Service: Amazon SageMaker Service

A directed edge connecting two lineage entities.

Contents

AssociationType

The type of the Association(Edge) between the source and destination. For example `ContributedTo`, `Produced`, or `DerivedFrom`.

Type: String

Valid Values: `ContributedTo | AssociatedWith | DerivedFrom | Produced | SameAs`

Required: No

DestinationArn

The Amazon Resource Name (ARN) of the destination lineage entity of the directed edge.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: No

SourceArn

The Amazon Resource Name (ARN) of the source lineage entity of the directed edge.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EdgeDeploymentConfig

Service: Amazon SageMaker Service

Contains information about the configuration of a deployment.

Contents

FailureHandlingPolicy

Toggle that determines whether to rollback to previous configuration if the current deployment fails. By default this is turned on. You may turn this off if you want to investigate the errors yourself.

Type: String

Valid Values: `ROLLBACK_ON_FAILURE | DO_NOTHING`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EdgeDeploymentModelConfig

Service: Amazon SageMaker Service

Contains information about the configuration of a model in a deployment.

Contents

EdgePackagingJobName

The edge packaging job associated with this deployment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelHandle

The name the device application uses to reference this model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EdgeDeploymentPlanSummary

Service: Amazon SageMaker Service

Contains information summarizing an edge deployment plan.

Contents

DeviceFleetName

The name of the device fleet used for the deployment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EdgeDeploymentFailed

The number of edge devices that failed the deployment.

Type: Integer

Required: Yes

EdgeDeploymentPending

The number of edge devices yet to pick up the deployment, or in progress.

Type: Integer

Required: Yes

EdgeDeploymentPlanArn

The ARN of the edge deployment plan.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z\-]*:\d{12}:edge-deployment/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

EdgeDeploymentPlanName

The name of the edge deployment plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EdgeDeploymentSuccess

The number of edge devices with the successful deployment.

Type: Integer

Required: Yes

CreationTime

The time when the edge deployment plan was created.

Type: Timestamp

Required: No

LastModifiedTime

The time when the edge deployment plan was last updated.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

EdgeDeploymentStatus

Service: Amazon SageMaker Service

Contains information summarizing the deployment stage results.

Contents

EdgeDeploymentFailedInStage

The number of edge devices that failed the deployment in current stage.

Type: Integer

Required: Yes

EdgeDeploymentPendingInStage

The number of edge devices yet to pick up the deployment in current stage, or in progress.

Type: Integer

Required: Yes

EdgeDeploymentSuccessInStage

The number of edge devices with the successful deployment in the current stage.

Type: Integer

Required: Yes

StageStatus

The general status of the current stage.

Type: String

Valid Values: `CREATING | READYTODEPLOY | STARTING | INPROGRESS | DEPLOYED | FAILED | STOPPING | STOPPED`

Required: Yes

EdgeDeploymentStageStartTime

The time when the deployment API started.

Type: Timestamp

Required: No

EdgeDeploymentStatusMessage

A detailed message about deployment status in current stage.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EdgeModel

Service: Amazon SageMaker Service

The model on the edge device.

Contents

ModelName

The name of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelVersion

The model version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 30.

Pattern: `[a-zA-Z0-9\ \_\.]+`

Required: Yes

LatestInference

The timestamp of the last inference that was made.

Type: Timestamp

Required: No

LatestSampleTime

The timestamp of the last data sample taken.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EdgeModelStat

Service: Amazon SageMaker Service

Status of edge devices with this model.

Contents

ActiveDeviceCount

The number of devices that have this model version, a heart beat, and are currently running.

Type: Long

Required: Yes

ConnectedDeviceCount

The number of devices that have this model version and have a heart beat.

Type: Long

Required: Yes

ModelName

The name of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelVersion

The model version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 30.

Pattern: `[a-zA-Z0-9\ \_\.]+`

Required: Yes

OfflineDeviceCount

The number of devices that have this model version and do not have a heart beat.

Type: Long

Required: Yes

SamplingDeviceCount

The number of devices with this model version and are producing sample data.

Type: Long

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EdgeModelSummary

Service: Amazon SageMaker Service

Summary of model on edge device.

Contents

ModelName

The name of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelVersion

The version model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 30.

Pattern: `[a-zA-Z0-9\ \_\.]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EdgeOutputConfig

Service: Amazon SageMaker Service

The output configuration.

Contents

S3OutputLocation

The Amazon Simple Storage (S3) bucker URI.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

KmsKeyId

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt data on the storage volume after compilation job. If you don't provide a KMS key ID, Amazon SageMaker uses the default KMS key for Amazon S3 for your role's account.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

PresetDeploymentConfig

The configuration used to create deployment artifacts. Specify configuration options with a JSON string. The available configuration options for each type are:

- `ComponentName` (optional) - Name of the GreenGrass V2 component. If not specified, the default name generated consists of "SagemakerEdgeManager" and the name of your SageMaker Edge Manager packaging job.
- `ComponentDescription` (optional) - Description of the component.
- `ComponentVersion` (optional) - The version of the component.

> **Note**
>
> AWS IoT Greengrass uses semantic versions for components. Semantic versions follow a *major.minor.patch* number system. For example, version 1.0.0 represents the first major release for a component. For more information, see the semantic version specification.

- `PlatformOS` (optional) - The name of the operating system for the platform. Supported platforms include Windows and Linux.
- `PlatformArchitecture` (optional) - The processor architecture for the platform.

  Supported architectures Windows include: Windows32_x86, Windows64_x64.

  Supported architectures for Linux include: Linux x86_64, Linux ARMV8.

Type: String

Required: No

PresetDeploymentType

The deployment type SageMaker Edge Manager will create. Currently only supports AWS IoT Greengrass Version 2 components.

Type: String

Valid Values: `GreengrassV2Component`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EdgePackagingJobSummary

Service: Amazon SageMaker Service

Summary of edge packaging job.

Contents

EdgePackagingJobArn

The Amazon Resource Name (ARN) of the edge packaging job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z\-]*:\d{12}:edge-packaging-job/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

EdgePackagingJobName

The name of the edge packaging job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EdgePackagingJobStatus

The status of the edge packaging job.

Type: String

Valid Values: `STARTING | INPROGRESS | COMPLETED | FAILED | STOPPING | STOPPED`

Required: Yes

CompilationJobName

The name of the SageMaker Neo compilation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

CreationTime

The timestamp of when the job was created.

Type: Timestamp

Required: No

LastModifiedTime

The timestamp of when the edge packaging job was last updated.

Type: Timestamp

Required: No

ModelName

The name of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelVersion

The version of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 30.

Pattern: `[a-zA-Z0-9\ \_\.]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EdgePresetDeploymentOutput

Service: Amazon SageMaker Service

The output of a SageMaker Edge Manager deployable resource.

Contents

Type

The deployment type created by SageMaker Edge Manager. Currently only supports AWS IoT Greengrass Version 2 components.

Type: String

Valid Values: `GreengrassV2Component`

Required: Yes

Artifact

The Amazon Resource Name (ARN) of the generated deployable resource.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Required: No

Status

The status of the deployable resource.

Type: String

Valid Values: `COMPLETED | FAILED`

Required: No

StatusMessage

Returns a message describing the status of the deployed resource.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EFSFileSystem

Service: Amazon SageMaker Service

A file system, created by you in Amazon EFS, that you assign to a user profile or space for an Amazon SageMaker AI Domain. Permitted users can access this file system in Amazon SageMaker AI Studio.

Contents

FileSystemId

The ID of your Amazon EFS file system.

Type: String

Length Constraints: Minimum length of 11. Maximum length of 21.

Pattern: `(fs-[0-9a-f]{8,})`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EFSFileSystemConfig

Service: Amazon SageMaker Service

The settings for assigning a custom Amazon EFS file system to a user profile or space for an Amazon SageMaker AI Domain.

Contents

FileSystemId

The ID of your Amazon EFS file system.

Type: String

Length Constraints: Minimum length of 11. Maximum length of 21.

Pattern: `(fs-[0-9a-f]{8,})`

Required: Yes

FileSystemPath

The path to the file system directory that is accessible in Amazon SageMaker AI Studio. Permitted users can access only this directory and below.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: \/\S*

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EmrServerlessComputeConfig

Service: Amazon SageMaker Service

> **Note**
>
> This data type is intended for use exclusively by SageMaker Canvas and cannot be used in other contexts at the moment.

Specifies the compute configuration for the EMR Serverless job.

Contents

ExecutionRoleARN

The ARN of the IAM role granting the AutoML job V2 the necessary permissions access policies to list, connect to, or manage EMR Serverless jobs. For detailed information about the required permissions of this role, see "How to configure AutoML to initiate a remote job on EMR Serverless for large datasets" in Create a regression or classification job for tabular data using the AutoML API or Create an AutoML job for time-series forecasting using the API.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EmrServerlessSettings

Service: Amazon SageMaker Service

The settings for running Amazon EMR Serverless jobs in SageMaker Canvas.

Contents

ExecutionRoleArn

The Amazon Resource Name (ARN) of the AWS IAM role that is assumed for running Amazon EMR Serverless jobs in SageMaker Canvas. This role should have the necessary permissions to read and write data attached and a trust relationship with EMR Serverless.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

Status

Describes whether Amazon EMR Serverless job capabilities are enabled or disabled in the SageMaker Canvas application.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EmrSettings

Service: Amazon SageMaker Service

The configuration parameters that specify the IAM roles assumed by the execution role of SageMaker (assumable roles) and the cluster instances or job execution environments (execution roles or runtime roles) to manage and access resources required for running Amazon EMR clusters or Amazon EMR Serverless applications.

Contents

AssumableRoleArns

An array of Amazon Resource Names (ARNs) of the IAM roles that the execution role of SageMaker can assume for performing operations or tasks related to Amazon EMR clusters or Amazon EMR Serverless applications. These roles define the permissions and access policies required when performing Amazon EMR-related operations, such as listing, connecting to, or terminating Amazon EMR clusters or Amazon EMR Serverless applications. They are typically used in cross-account access scenarios, where the Amazon EMR resources (clusters or serverless applications) are located in a different AWS account than the SageMaker domain.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

ExecutionRoleArns

An array of Amazon Resource Names (ARNs) of the IAM roles used by the Amazon EMR cluster instances or job execution environments to access other AWS services and resources needed during the runtime of your Amazon EMR or Amazon EMR Serverless workloads, such as Amazon S3 for data access, Amazon CloudWatch for logging, or other AWS services based on the particular workload requirements.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EMRStepMetadata

Service: Amazon SageMaker Service

The configurations and outcomes of an Amazon EMR step execution.

Contents

ClusterId

The identifier of the EMR cluster.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

LogFilePath

The path to the log file where the cluster step's failure root cause is recorded.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

StepId

The identifier of the EMR cluster step.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

StepName

The name of the EMR cluster step.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Endpoint

Service: Amazon SageMaker Service

A hosted endpoint for real-time inference.

Contents

CreationTime

The time that the endpoint was created.

Type: Timestamp

Required: Yes

EndpointArn

The Amazon Resource Name (ARN) of the endpoint.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint/.*`

Required: Yes

EndpointConfigName

The endpoint configuration associated with the endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EndpointName

The name of the endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EndpointStatus

The status of the endpoint.

Type: String

Valid Values: `OutOfService | Creating | Updating | SystemUpdating | RollingBack | InService | Deleting | Failed | UpdateRollbackFailed`

Required: Yes

LastModifiedTime

The last time the endpoint was modified.

Type: Timestamp

Required: Yes

DataCaptureConfig

The currently active data capture configuration used by your Endpoint.

Type: DataCaptureConfigSummary object

Required: No

FailureReason

If the endpoint failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

MonitoringSchedules

A list of monitoring schedules for the endpoint. For information about model monitoring, see Amazon SageMaker Model Monitor.

Type: Array of MonitoringSchedule objects

Required: No

ProductionVariants

A list of the production variants hosted on the endpoint. Each production variant is a model.

Type: Array of ProductionVariantSummary objects

Array Members: Minimum number of 1 item.

Required: No

ShadowProductionVariants

A list of the shadow variants hosted on the endpoint. Each shadow variant is a model in shadow mode with production traffic replicated from the production variant.

Type: Array of ProductionVariantSummary objects

Array Members: Minimum number of 1 item.

Required: No

Tags

A list of the tags associated with the endpoint. For more information, see Tagging AWS resources in the *AWS General Reference Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EndpointConfigStepMetadata

Service: Amazon SageMaker Service

Metadata for an endpoint configuration step.

Contents

Arn

The Amazon Resource Name (ARN) of the endpoint configuration used in the step.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint-config/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EndpointConfigSummary

Service: Amazon SageMaker Service

Provides summary information for an endpoint configuration.

Contents

CreationTime

A timestamp that shows when the endpoint configuration was created.

Type: Timestamp

Required: Yes

EndpointConfigArn

The Amazon Resource Name (ARN) of the endpoint configuration.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint-config/.*`

Required: Yes

EndpointConfigName

The name of the endpoint configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EndpointInfo

Service: Amazon SageMaker Service

Details about a customer endpoint that was compared in an Inference Recommender job.

Contents

EndpointName

The name of a customer's endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EndpointInput

Service: Amazon SageMaker Service

Input object for the endpoint

Contents

EndpointName

An endpoint in customer's account which has enabled `DataCaptureConfig` enabled.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

LocalPath

Path to the filesystem where the endpoint data is available to the container.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

EndTimeOffset

If specified, monitoring jobs substract this time from the end time. For information about using offsets for scheduling monitoring jobs, see Schedule Model Quality Monitoring Jobs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 15.

Pattern: `.?P.*`

Required: No

ExcludeFeaturesAttribute

The attributes of the input data to exclude from the analysis.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 100.

Required: No

FeaturesAttribute

The attributes of the input data that are the input features.

Type: String

Required: No

InferenceAttribute

The attribute of the input data that represents the ground truth label.

Type: String

Required: No

ProbabilityAttribute

In a classification problem, the attribute that represents the class probability.

Type: String

Required: No

ProbabilityThresholdAttribute

The threshold for the class probability to be evaluated as a positive result.

Type: Double

Required: No

S3DataDistributionType

Whether input data distributed in Amazon S3 is fully replicated or sharded by an Amazon S3 key. Defaults to `FullyReplicated`

Type: String

Valid Values: `FullyReplicated | ShardedByS3Key`

Required: No

S3InputMode

Whether the `Pipe` or `File` is used as the input mode for transferring data for the monitoring job. `Pipe` mode is recommended for large datasets. `File` mode is useful for small files that fit in memory. Defaults to `File`.

Type: String

Valid Values: `Pipe | File`

Required: No

StartTimeOffset

If specified, monitoring jobs substract this time from the start time. For information about using offsets for scheduling monitoring jobs, see Schedule Model Quality Monitoring Jobs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 15.

Pattern: `.?P.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EndpointInputConfiguration

Service: Amazon SageMaker Service

The endpoint configuration for the load test.

Contents

EnvironmentParameterRanges

The parameter you want to benchmark against.

Type: EnvironmentParameterRanges object

Required: No

InferenceSpecificationName

The inference specification name in the model package version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

InstanceType

The instance types to use for the load test.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge |`

ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large | ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge | ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large | ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge | ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge | ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge | ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge | ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge | ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge | ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large | ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge | ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge | ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge | ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge |

```
ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge
| ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge |
ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge
| ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge
| ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge |
ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge
| ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge |
ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge
| ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge |
ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge |
ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium
| ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge |
ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge
| ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge |
ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge
| ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge |
ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge |
ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large |
ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge
| ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge |
ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-
gb200.36xlarge | ml.p5.4xlarge
```

Required: No

ServerlessConfig

Specifies the serverless configuration for an endpoint variant.

Type: ProductionVariantServerlessConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

EndpointMetadata

Service: Amazon SageMaker Service

The metadata of the endpoint.

Contents

EndpointName

The name of the endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EndpointConfigName

The name of the endpoint configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

EndpointStatus

The status of the endpoint. For possible values of the status of an endpoint, see EndpointSummary.

Type: String

Valid Values: `OutOfService | Creating | Updating | SystemUpdating | RollingBack | InService | Deleting | Failed | UpdateRollbackFailed`

Required: No

FailureReason

If the status of the endpoint is `Failed`, or the status is `InService` but update operation fails, this provides the reason why it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EndpointOutputConfiguration

Service: Amazon SageMaker Service

The endpoint configuration made by Inference Recommender during a recommendation job.

Contents

EndpointName

The name of the endpoint made during a recommendation job.

Type: String

Required: Yes

VariantName

The name of the production variant (deployed model) made during a recommendation job.

Type: String

Required: Yes

InitialInstanceCount

The number of instances recommended to launch initially.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

InstanceType

The instance type recommended by Amazon SageMaker Inference Recommender.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge |`

ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge
| ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge |
ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge |
ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge |
ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge
| ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge
| ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge
| ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge |
ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge
| ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large
| ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge |
ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge
| ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge |
ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge
| ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge |
ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge
| ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge
| ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge
| ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge
| ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge
| ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large |
ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge |
ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge
| ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge |
ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge
| ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge
| ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large
| ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge
| ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge |
ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge |
ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge |
ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge
| ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge |
ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large |
ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge |
ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge |

ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge | ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge | ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium | ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge | ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge | ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge | ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge | ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge | ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge | ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large | ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge | ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge | ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge

Required: No

ServerlessConfig

Specifies the serverless configuration for an endpoint variant.

Type: ProductionVariantServerlessConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EndpointPerformance

Service: Amazon SageMaker Service

The performance results from running an Inference Recommender job on an existing endpoint.

Contents

EndpointInfo

Details about a customer endpoint that was compared in an Inference Recommender job.

Type: EndpointInfo object

Required: Yes

Metrics

The metrics for an existing endpoint.

Type: InferenceMetrics object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EndpointStepMetadata

Service: Amazon SageMaker Service

Metadata for an endpoint step.

Contents

Arn

The Amazon Resource Name (ARN) of the endpoint in the step.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EndpointSummary

Service: Amazon SageMaker Service

Provides summary information for an endpoint.

Contents

CreationTime

A timestamp that shows when the endpoint was created.

Type: Timestamp

Required: Yes

EndpointArn

The Amazon Resource Name (ARN) of the endpoint.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint/.*`

Required: Yes

EndpointName

The name of the endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EndpointStatus

The status of the endpoint.

- `OutOfService`: Endpoint is not available to take incoming requests.
- `Creating`: CreateEndpoint is executing.

- `Updating`: UpdateEndpoint or UpdateEndpointWeightsAndCapacities is executing.
- `SystemUpdating`: Endpoint is undergoing maintenance and cannot be updated or deleted or re-scaled until it has completed. This maintenance operation does not change any customer-specified values such as VPC config, KMS encryption, model, instance type, or instance count.
- `RollingBack`: Endpoint fails to scale up or down or change its variant weight and is in the process of rolling back to its previous configuration. Once the rollback completes, endpoint returns to an `InService` status. This transitional status only applies to an endpoint that has autoscaling enabled and is undergoing variant weight or capacity changes as part of an UpdateEndpointWeightsAndCapacities call or when the UpdateEndpointWeightsAndCapacities operation is called explicitly.
- `InService`: Endpoint is available to process incoming requests.
- `Deleting`: DeleteEndpoint is executing.
- `Failed`: Endpoint could not be created, updated, or re-scaled. Use `DescribeEndpointOutput$FailureReason` for information about the failure. DeleteEndpoint is the only operation that can be performed on a failed endpoint.

To get a list of endpoints with a specified status, use the `StatusEquals` filter with a call to ListEndpoints.

Type: String

Valid Values: `OutOfService | Creating | Updating | SystemUpdating | RollingBack | InService | Deleting | Failed | UpdateRollbackFailed`

Required: Yes

LastModifiedTime

A timestamp that shows when the endpoint was last modified.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EnvironmentConfig

Service: Amazon SageMaker Service

The configuration for the restricted instance groups (RIG) environment.

Contents

FSxLustreConfig

Configuration settings for an Amazon FSx for Lustre file system to be used with the cluster.

Type: FSxLustreConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EnvironmentConfigDetails

Service: Amazon SageMaker Service

The configuration details for the restricted instance groups (RIG) environment.

Contents

FSxLustreConfig

Configuration settings for an Amazon FSx for Lustre file system to be used with the cluster.

Type: FSxLustreConfig object

Required: No

S3OutputPath

The Amazon S3 path where output data from the restricted instance group (RIG) environment will be stored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EnvironmentParameter

Service: Amazon SageMaker Service

A list of environment parameters suggested by the Amazon SageMaker Inference Recommender.

Contents

Key

The environment key suggested by the Amazon SageMaker Inference Recommender.

Type: String

Required: Yes

Value

The value suggested by the Amazon SageMaker Inference Recommender.

Type: String

Required: Yes

ValueType

The value type suggested by the Amazon SageMaker Inference Recommender.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EnvironmentParameterRanges

Service: Amazon SageMaker Service

Specifies the range of environment parameters

Contents

CategoricalParameterRanges

Specified a list of parameters for each category.

Type: Array of CategoricalParameter objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ErrorInfo

Service: Amazon SageMaker Service

This is an error field object that contains the error code and the reason for an operation failure.

Contents

Code

The error code for an invalid or failed operation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `(?!\s*$).+`

Required: No

Reason

The failure reason for the operation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(?!\s*$).+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EventDetails

Service: Amazon SageMaker Service

Detailed information about a specific event, including event metadata.

Contents

EventMetadata

Metadata specific to the event, which may include information about the cluster, instance group, or instance involved.

Type: EventMetadata object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EventMetadata

Service: Amazon SageMaker Service

Metadata associated with a cluster event, which may include details about various resource types.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

Cluster

Metadata specific to cluster-level events.

Type: ClusterMetadata object

Required: No

Instance

Metadata specific to instance-level events.

Type: InstanceMetadata object

Required: No

InstanceGroup

Metadata specific to instance group-level events.

Type: InstanceGroupMetadata object

Required: No

InstanceGroupScaling

Metadata related to instance group scaling events.

Type: InstanceGroupScalingMetadata object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Experiment

Service: Amazon SageMaker Service

The properties of an experiment as returned by the Search API. For information about experiments, see the CreateExperiment API.

Contents

CreatedBy

Who created the experiment.

Type: UserContext object

Required: No

CreationTime

When the experiment was created.

Type: Timestamp

Required: No

Description

The description of the experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

DisplayName

The name of the experiment as displayed. If `DisplayName` isn't specified, `ExperimentName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

ExperimentArn

The Amazon Resource Name (ARN) of the experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment/.*`

Required: No

ExperimentName

The name of the experiment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

LastModifiedTime

When the experiment was last modified.

Type: Timestamp

Required: No

Source

The source of the experiment.

Type: ExperimentSource object

Required: No

Tags

The list of tags that are associated with the experiment. You can use Search API to search on the tags.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExperimentConfig

Service: Amazon SageMaker Service

Associates a SageMaker job as a trial component with an experiment and trial. Specified when you call the following APIs:

- CreateProcessingJob
- CreateTrainingJob
- CreateTransformJob

Contents

ExperimentName

The name of an existing experiment to associate with the trial component.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

RunName

The name of the experiment run to associate with the trial component.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

TrialComponentDisplayName

The display name for the trial component. If this key isn't specified, the display name is the trial component name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

TrialName

The name of an existing trial to associate the trial component with. If not specified, a new trial is created.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExperimentSource

Service: Amazon SageMaker Service

The source of the experiment.

Contents

SourceArn

The Amazon Resource Name (ARN) of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:.*`

Required: Yes

SourceType

The source type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExperimentSummary

Service: Amazon SageMaker Service

A summary of the properties of an experiment. To get the complete set of properties, call the DescribeExperiment API and provide the `ExperimentName`.

Contents

CreationTime

When the experiment was created.

Type: Timestamp

Required: No

DisplayName

The name of the experiment as displayed. If `DisplayName` isn't specified, `ExperimentName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

ExperimentArn

The Amazon Resource Name (ARN) of the experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment/.*`

Required: No

ExperimentName

The name of the experiment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

ExperimentSource

The source of the experiment.

Type: ExperimentSource object

Required: No

LastModifiedTime

When the experiment was last modified.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Explainability

Service: Amazon SageMaker Service

Contains explainability metrics for a model.

Contents

Report

The explainability report for a model.

Type: MetricsSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExplainerConfig

Service: Amazon SageMaker Service

A parameter to activate explainers.

Contents

ClarifyExplainerConfig

A member of `ExplainerConfig` that contains configuration parameters for the SageMaker Clarify explainer.

Type: ClarifyExplainerConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FailStepMetadata

Service: Amazon SageMaker Service

The container for the metadata for Fail step.

Contents

ErrorMessage

A message that you define and then is processed and rendered by the Fail step when the error occurs.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FeatureDefinition

Service: Amazon SageMaker Service

A list of features. You must include `FeatureName` and `FeatureType`. Valid feature `FeatureTypes` are `Integral`, `Fractional` and `String`.

Contents

FeatureName

The name of a feature. The type must be a string. `FeatureName` cannot be any of the following: `is_deleted`, `write_time`, `api_invocation_time`.

The name:

- Must start with an alphanumeric character.
- Can only include alphanumeric characters, underscores, and hyphens. Spaces are not allowed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

Required: Yes

FeatureType

The value type of a feature. Valid values are Integral, Fractional, or String.

Type: String

Valid Values: `Integral | Fractional | String`

Required: Yes

CollectionConfig

Configuration for your collection.

Type: CollectionConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

CollectionType

A grouping of elements where each element within the collection must have the same feature type (`String`, `Integral`, or `Fractional`).

- `List`: An ordered collection of elements.
- `Set`: An unordered collection of unique elements.
- `Vector`: A specialized list that represents a fixed-size array of elements. The vector dimension is determined by you. Must have elements with fractional feature types.

Type: String

Valid Values: `List | Set | Vector`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FeatureGroup

Service: Amazon SageMaker Service

Amazon SageMaker Feature Store stores features in a collection called Feature Group. A Feature Group can be visualized as a table which has rows, with a unique identifier for each row where each column in the table is a feature. In principle, a Feature Group is composed of features and values per features.

Contents

CreationTime

The time a `FeatureGroup` was created.

Type: Timestamp

Required: No

Description

A free form description of a `FeatureGroup`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Required: No

EventTimeFeatureName

The name of the feature that stores the `EventTime` of a Record in a `FeatureGroup`.

A `EventTime` is point in time when a new event occurs that corresponds to the creation or update of a `Record` in `FeatureGroup`. All `Records` in the `FeatureGroup` must have a corresponding `EventTime`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

Required: No

FailureReason

The reason that the `FeatureGroup` failed to be replicated in the `OfflineStore`. This is failure may be due to a failure to create a `FeatureGroup` in or delete a `FeatureGroup` from the `OfflineStore`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

FeatureDefinitions

A list of `Features`. Each `Feature` must include a `FeatureName` and a `FeatureType`.

Valid `FeatureTypes` are `Integral`, `Fractional` and `String`.

`FeatureNames` cannot be any of the following: `is_deleted`, `write_time`, `api_invocation_time`.

You can create up to 2,500 `FeatureDefinitions` per `FeatureGroup`.

Type: Array of FeatureDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 2500 items.

Required: No

FeatureGroupArn

The Amazon Resource Name (ARN) of a `FeatureGroup`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group/.*`

Required: No

FeatureGroupName

The name of the `FeatureGroup`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63}`

Required: No

FeatureGroupStatus

A `FeatureGroup` status.

Type: String

Valid Values: `Creating | Created | CreateFailed | Deleting | DeleteFailed`

Required: No

LastModifiedTime

A timestamp indicating the last time you updated the feature group.

Type: Timestamp

Required: No

LastUpdateStatus

A value that indicates whether the feature group was updated successfully.

Type: LastUpdateStatus object

Required: No

OfflineStoreConfig

The configuration of an `OfflineStore`.

Provide an `OfflineStoreConfig` in a request to `CreateFeatureGroup` to create an `OfflineStore`.

To encrypt an `OfflineStore` using at rest data encryption, specify AWS Key Management Service (KMS) key ID, or `KMSKeyId`, in `S3StorageConfig`.

Type: OfflineStoreConfig object

Required: No

OfflineStoreStatus

The status of `OfflineStore`.

Type: OfflineStoreStatus object

Required: No

OnlineStoreConfig

Use this to specify the AWS Key Management Service (KMS) Key ID, or `KMSKeyId`, for at rest data encryption. You can turn `OnlineStore` on or off by specifying the `EnableOnlineStore` flag at General Assembly.

The default value is `False`.

Type: OnlineStoreConfig object

Required: No

RecordIdentifierFeatureName

The name of the `Feature` whose value uniquely identifies a `Record` defined in the `FeatureGroup` `FeatureDefinitions`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

Required: No

RoleArn

The Amazon Resource Name (ARN) of the IAM execution role used to create the feature group.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

Tags

Tags used to define a `FeatureGroup`.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FeatureGroupSummary

Service: Amazon SageMaker Service

The name, ARN, `CreationTime`, `FeatureGroup` values, `LastUpdatedTime` and `EnableOnlineStorage` status of a `FeatureGroup`.

Contents

CreationTime

A timestamp indicating the time of creation time of the `FeatureGroup`.

Type: Timestamp

Required: Yes

FeatureGroupArn

Unique identifier for the `FeatureGroup`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group/.*`

Required: Yes

FeatureGroupName

The name of `FeatureGroup`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63}`

Required: Yes

FeatureGroupStatus

The status of a FeatureGroup. The status can be any of the following: `Creating`, `Created`, `CreateFail`, `Deleting` or `DetailFail`.

Type: String

Valid Values: `Creating | Created | CreateFailed | Deleting | DeleteFailed`

Required: No

OfflineStoreStatus

Notifies you if replicating data into the `OfflineStore` has failed. Returns either: `Active` or `Blocked`.

Type: OfflineStoreStatus object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FeatureMetadata

Service: Amazon SageMaker Service

The metadata for a feature. It can either be metadata that you specify, or metadata that is updated automatically.

Contents

CreationTime

A timestamp indicating when the feature was created.

Type: Timestamp

Required: No

Description

An optional description that you specify to better describe the feature.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

FeatureGroupArn

The Amazon Resource Number (ARN) of the feature group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group/.*`

Required: No

FeatureGroupName

The name of the feature group containing the feature.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63}`

Required: No

FeatureName

The name of feature.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

Required: No

FeatureType

The data type of the feature.

Type: String

Valid Values: `Integral | Fractional | String`

Required: No

LastModifiedTime

A timestamp indicating when the feature was last modified.

Type: Timestamp

Required: No

Parameters

Optional key-value pairs that you specify to better describe the feature.

Type: Array of FeatureParameter objects

Array Members: Minimum number of 0 items. Maximum number of 25 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FeatureParameter

Service: Amazon SageMaker Service

A key-value pair that you specify to describe the feature.

Contents

Key

A key that must contain a value to describe the feature.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `([\p{L}\p{Z}\p{N}_.:/=+\-]*)`

Required: No

Value

The value that belongs to a key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `([\p{L}\p{Z}\p{N}_.:/=+\-]*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FileSource

Service: Amazon SageMaker Service

Contains details regarding the file source.

Contents

S3Uri

The Amazon S3 URI for the file source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

ContentDigest

The digest of the file source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 72.

Pattern: `[Ss][Hh][Aa]256:[0-9a-fA-F]{64}`

Required: No

ContentType

The type of content stored in the file source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FileSystemConfig

Service: Amazon SageMaker Service

The Amazon Elastic File System storage configuration for a SageMaker AI image.

Contents

DefaultGid

The default POSIX group ID (GID). If not specified, defaults to `100`.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 65535.

Required: No

DefaultUid

The default POSIX user ID (UID). If not specified, defaults to `1000`.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 65535.

Required: No

MountPath

The path within the image to mount the user's EFS home directory. The directory should be empty. If not specified, defaults to */home/sagemaker-user*.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `\/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FileSystemDataSource

Service: Amazon SageMaker Service

Specifies a file system data source for a channel.

Contents

DirectoryPath

The full path to the directory to associate with the channel.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4096.

Pattern: `.*`

Required: Yes

FileSystemAccessMode

The access mode of the mount of the directory associated with the channel. A directory can be mounted either in `ro` (read-only) or `rw` (read-write) mode.

Type: String

Valid Values: `rw | ro`

Required: Yes

FileSystemId

The file system id.

Type: String

Length Constraints: Minimum length of 11. Maximum length of 21.

Pattern: `(fs-[0-9a-f]{8,})`

Required: Yes

FileSystemType

The file system type.

Type: String

Valid Values: `EFS | FSxLustre`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Filter

Service: Amazon SageMaker Service

A conditional statement for a search expression that includes a resource property, a Boolean operator, and a value. Resources that match the statement are returned in the results from the Search API.

If you specify a `Value`, but not an `Operator`, SageMaker uses the equals operator.

In search, there are several property types:

Metrics

To define a metric filter, enter a value using the form `"Metrics.<name>"`, where `<name>` is a metric name. For example, the following filter searches for training jobs with an `"accuracy"` metric greater than `"0.9"`:

```
{
"Name": "Metrics.accuracy",
"Operator": "GreaterThan",
"Value": "0.9"
}
```

HyperParameters

To define a hyperparameter filter, enter a value with the form `"HyperParameters.<name>"`. Decimal hyperparameter values are treated as a decimal in a comparison if the specified `Value` is also a decimal value. If the specified `Value` is an integer, the decimal hyperparameter values are treated as integers. For example, the following filter is satisfied by training jobs with a `"learning_rate"` hyperparameter that is less than `"0.5"`:

```
{
"Name": "HyperParameters.learning_rate",
"Operator": "LessThan",
"Value": "0.5"
```

}

Tags

To define a tag filter, enter a value with the form `Tags.<key>`.

Contents

Name

A resource property name. For example, `TrainingJobName`. For valid property names, see SearchRecord. You must specify a valid property for the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: Yes

Operator

A Boolean binary operator that is used to evaluate the filter. The operator field contains one of the following values:

Equals

The value of `Name` equals `Value`.

NotEquals

The value of `Name` doesn't equal `Value`.

Exists

The `Name` property exists.

NotExists

The `Name` property does not exist.

GreaterThan

The value of `Name` is greater than `Value`. Not supported for text properties.

GreaterThanOrEqualTo

The value of `Name` is greater than or equal to `Value`. Not supported for text properties.

LessThan

The value of `Name` is less than `Value`. Not supported for text properties.

LessThanOrEqualTo

The value of `Name` is less than or equal to `Value`. Not supported for text properties.

In

The value of `Name` is one of the comma delimited strings in `Value`. Only supported for text properties.

Contains

The value of `Name` contains the string `Value`. Only supported for text properties.

A `SearchExpression` can include the `Contains` operator multiple times when the value of `Name` is one of the following:

- `Experiment.DisplayName`
- `Experiment.ExperimentName`
- `Experiment.Tags`
- `Trial.DisplayName`
- `Trial.TrialName`
- `Trial.Tags`
- `TrialComponent.DisplayName`
- `TrialComponent.TrialComponentName`
- `TrialComponent.Tags`
- `TrialComponent.InputArtifacts`
- `TrialComponent.OutputArtifacts`

A `SearchExpression` can include only one `Contains` operator for all other values of `Name`. In these cases, if you include multiple `Contains` operators in the `SearchExpression`, the result is the following error message: `"'CONTAINS' operator usage limit of 1 exceeded."`

Type: String

Valid Values: `Equals | NotEquals | GreaterThan | GreaterThanOrEqualTo | LessThan | LessThanOrEqualTo | Contains | Exists | NotExists | In`

Required: No

Value

A value used with `Name` and `Operator` to determine which resources satisfy the filter's condition. For numerical properties, `Value` must be an integer or floating-point decimal. For timestamp properties, `Value` must be an ISO 8601 date-time string of the following format: `YYYY-mm-dd'T'HH:MM:SS`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `.+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FinalAutoMLJobObjectiveMetric

Service: Amazon SageMaker Service

The best candidate result from an AutoML training job.

Contents

MetricName

The name of the metric with the best result. For a description of the possible objective metrics, see AutoMLJobObjective$MetricName.

Type: String

Valid Values: `Accuracy | MSE | F1 | F1macro | AUC | RMSE | BalancedAccuracy | R2 | Recall | RecallMacro | Precision | PrecisionMacro | MAE | MAPE | MASE | WAPE | AverageWeightedQuantileLoss`

Required: Yes

Value

The value of the metric with the best result.

Type: Float

Required: Yes

StandardMetricName

The name of the standard metric. For a description of the standard metrics, see Autopilot candidate metrics.

Type: String

Valid Values: `Accuracy | MSE | F1 | F1macro | AUC | RMSE | BalancedAccuracy | R2 | Recall | RecallMacro | Precision | PrecisionMacro | MAE | MAPE | MASE | WAPE | AverageWeightedQuantileLoss`

Required: No

Type

The type of metric with the best result.

Type: String

Valid Values: `Maximize | Minimize`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FinalHyperParameterTuningJobObjectiveMetric

Service: Amazon SageMaker Service

Shows the latest objective metric emitted by a training job that was launched by a hyperparameter tuning job. You define the objective metric in the `HyperParameterTuningJobObjective` parameter of HyperParameterTuningJobConfig.

Contents

MetricName

The name of the objective metric. For SageMaker built-in algorithms, metrics are defined per algorithm. See the metrics for XGBoost as an example. You can also use a custom algorithm for training and define your own metrics. For more information, see Define metrics and environment variables.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: Yes

Value

The value of the objective metric.

Type: Float

Required: Yes

Type

Select if you want to minimize or maximize the objective metric during hyperparameter tuning.

Type: String

Valid Values: `Maximize | Minimize`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowDefinitionOutputConfig

Service: Amazon SageMaker Service

Contains information about where human output will be stored.

Contents

S3OutputPath

The Amazon S3 path where the object containing human output will be made available.

To learn more about the format of Amazon A2I output data, see Amazon A2I Output Data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

KmsKeyId

The Amazon Key Management Service (KMS) key ID for server-side encryption.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowDefinitionSummary

Service: Amazon SageMaker Service

Contains summary information about the flow definition.

Contents

CreationTime

The timestamp when SageMaker created the flow definition.

Type: Timestamp

Required: Yes

FlowDefinitionArn

The Amazon Resource Name (ARN) of the flow definition.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]+:[0-9]{12}:flow-definition/.*`

Required: Yes

FlowDefinitionName

The name of the flow definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9]){0,62}`

Required: Yes

FlowDefinitionStatus

The status of the flow definition. Valid values:

Type: String

Valid Values: `Initializing | Active | Failed | Deleting`

Required: Yes

FailureReason

The reason why the flow definition creation failed. A failure reason is returned only when the flow definition status is `Failed`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FSxLustreConfig

Service: Amazon SageMaker Service

Configuration settings for an Amazon FSx for Lustre file system to be used with the cluster.

Contents

PerUnitStorageThroughput

The throughput capacity of the Amazon FSx for Lustre file system, measured in MB/s per TiB of storage.

Type: Integer

Valid Range: Minimum value of 125. Maximum value of 1000.

Required: Yes

SizeInGiB

The storage capacity of the Amazon FSx for Lustre file system, specified in gibibytes (GiB).

Type: Integer

Valid Range: Minimum value of 1200. Maximum value of 100800.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FSxLustreFileSystem

Service: Amazon SageMaker Service

A custom file system in Amazon FSx for Lustre.

Contents

FileSystemId

Amazon FSx for Lustre file system ID.

Type: String

Length Constraints: Minimum length of 11. Maximum length of 21.

Pattern: `(fs-[0-9a-f]{8,})`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FSxLustreFileSystemConfig

Service: Amazon SageMaker Service

The settings for assigning a custom Amazon FSx for Lustre file system to a user profile or space for an Amazon SageMaker Domain.

Contents

FileSystemId

The globally unique, 17-digit, ID of the file system, assigned by Amazon FSx for Lustre.

Type: String

Length Constraints: Minimum length of 11. Maximum length of 21.

Pattern: `(fs-[0-9a-f]{8,})`

Required: Yes

FileSystemPath

The path to the file system directory that is accessible in Amazon SageMaker Studio. Permitted users can access only this directory and below.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `\/\S*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GenerativeAiSettings

Service: Amazon SageMaker Service

The generative AI settings for the SageMaker Canvas application.

Configure these settings for Canvas users starting chats with generative AI foundation models. For more information, see Use generative AI with foundation models.

Contents

AmazonBedrockRoleArn

The ARN of an AWS IAM role that allows fine-tuning of large language models (LLMs) in Amazon Bedrock. The IAM role should have Amazon S3 read and write permissions, as well as a trust relationship that establishes `bedrock.amazonaws.com` as a service principal.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GitConfig

Service: Amazon SageMaker Service

Specifies configuration details for a Git repository in your AWS account.

Contents

RepositoryUrl

The URL where the Git repository is located.

Type: String

Length Constraints: Minimum length of 11. Maximum length of 1024.

Pattern: `https://([^/]+)/?.{3,1016}`

Required: Yes

Branch

The default branch for the Git repository.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[^ ~^:?*\[]+`

Required: No

SecretArn

The Amazon Resource Name (ARN) of the AWS Secrets Manager secret that contains the credentials used to access the git repository. The secret must have a staging label of `AWSCURRENT` and must be in the following format:

`{"username": *UserName*, "password": *Password*}`

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:secretsmanager:[a-z0-9\-]*:[0-9]{12}:secret:.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GitConfigForUpdate

Service: Amazon SageMaker Service

Specifies configuration details for a Git repository when the repository is updated.

Contents

SecretArn

The Amazon Resource Name (ARN) of the AWS Secrets Manager secret that contains the credentials used to access the git repository. The secret must have a staging label of `AWSCURRENT` and must be in the following format:

`{"username": UserName, "password": Password}`

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:secretsmanager:[a-z0-9\-]*:[0-9]{12}:secret:.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HiddenSageMakerImage

Service: Amazon SageMaker Service

The SageMaker images that are hidden from the Studio user interface. You must specify the SageMaker image name and version aliases.

Contents

SageMakerImageName

The SageMaker image name that you are hiding from the Studio user interface.

Type: String

Valid Values: `sagemaker_distribution`

Required: No

VersionAliases

The version aliases you are hiding from the Studio user interface.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(0|[1-9]\d*)\.(0|[1-9]\d*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HolidayConfigAttributes

Service: Amazon SageMaker Service

Stores the holiday featurization attributes applicable to each item of time-series datasets during the training of a forecasting model. This allows the model to identify patterns associated with specific holidays.

Contents

CountryCode

The country code for the holiday calendar.

For the list of public holiday calendars supported by AutoML job V2, see Country Codes. Use the country code corresponding to the country of your choice.

Type: String

Length Constraints: Fixed length of 2.

Pattern: `[A-Z]{2}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HubAccessConfig

Service: Amazon SageMaker Service

The configuration for a private hub model reference that points to a public SageMaker JumpStart model.

For more information about private hubs, see Private curated hubs for foundation model access control in JumpStart.

Contents

HubContentArn

The ARN of your private model hub content. This should be a `ModelReference` resource type that points to a SageMaker JumpStart public hub model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HubContentDependency

Service: Amazon SageMaker Service

Any dependencies related to hub content, such as scripts, model artifacts, datasets, or notebooks.

Contents

DependencyCopyPath

The hub content dependency copy path.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Pattern: `.*`

Required: No

DependencyOriginPath

The hub content dependency origin path.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HubContentInfo

Service: Amazon SageMaker Service

Information about hub content.

Contents

CreationTime

The date and time that the hub content was created.

Type: Timestamp

Required: Yes

DocumentSchemaVersion

The version of the hub content document schema.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: Yes

HubContentArn

The Amazon Resource Name (ARN) of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

HubContentName

The name of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubContentStatus

The status of the hub content.

Type: String

Valid Values: `Available | Importing | Deleting | ImportFailed | DeleteFailed | PendingImport | PendingDelete`

Required: Yes

HubContentType

The type of hub content.

Type: String

Valid Values: `Model | Notebook | ModelReference | DataSet | JsonDoc`

Required: Yes

HubContentVersion

The version of the hub content.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 14.

Pattern: `\d{1,4}.\d{1,4}.\d{1,4}`

Required: Yes

HubContentDescription

A description of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Pattern: `.*`

Required: No

HubContentDisplayName

The display name of the hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

HubContentSearchKeywords

The searchable keywords for the hub content.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

OriginalCreationTime

The date and time when the hub content was originally created, before any updates or revisions.

Type: Timestamp

Required: No

SageMakerPublicHubContentArn

The ARN of the public hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:aws:hub-content\/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}\/Model\/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: No

SupportStatus

The support status of the hub content.

Type: String

Valid Values: `Supported | Deprecated | Restricted`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HubInfo

Service: Amazon SageMaker Service

Information about a hub.

Contents

CreationTime

The date and time that the hub was created.

Type: Timestamp

Required: Yes

HubArn

The Amazon Resource Name (ARN) of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

HubName

The name of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

HubStatus

The status of the hub.

Type: String

Valid Values: `InService | Creating | Updating | Deleting | CreateFailed | UpdateFailed | DeleteFailed`

Required: Yes

LastModifiedTime

The date and time that the hub was last modified.

Type: Timestamp

Required: Yes

HubDescription

A description of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Pattern: `.*`

Required: No

HubDisplayName

The display name of the hub.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

HubSearchKeywords

The searchable keywords for the hub.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[^A-Z]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HubS3StorageConfig

Service: Amazon SageMaker Service

The Amazon S3 storage configuration of a hub.

Contents

S3OutputPath

The Amazon S3 bucket prefix for hosting hub content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HumanLoopActivationConditionsConfig

Service: Amazon SageMaker Service

Defines under what conditions SageMaker creates a human loop. Used within CreateFlowDefinition. See HumanLoopActivationConditionsConfig for the required format of activation conditions.

Contents

HumanLoopActivationConditions

JSON expressing use-case specific conditions declaratively. If any condition is matched, atomic tasks are created against the configured work team. The set of conditions is different for Rekognition and Textract. For more information about how to structure the JSON, see JSON Schema for Human Loop Activation Conditions in Amazon Augmented AI in the *Amazon SageMaker Developer Guide*.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10240.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HumanLoopActivationConfig

Service: Amazon SageMaker Service

Provides information about how and under what conditions SageMaker creates a human loop. If `HumanLoopActivationConfig` is not given, then all requests go to humans.

Contents

HumanLoopActivationConditionsConfig

Container structure for defining under what conditions SageMaker creates a human loop.

Type: HumanLoopActivationConditionsConfig object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HumanLoopConfig

Service: Amazon SageMaker Service

Describes the work to be performed by human workers.

Contents

HumanTaskUiArn

The Amazon Resource Name (ARN) of the human task user interface.

You can use standard HTML and Crowd HTML Elements to create a custom worker task template. You use this template to create a human task UI.

To learn how to create a custom HTML template, see Create Custom Worker Task Template.

To learn how to create a human task UI, which is a worker task template that can be used in a flow definition, see Create and Delete a Worker Task Templates.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]+:[0-9]{12}:human-task-ui/.*`

Required: Yes

TaskCount

The number of distinct workers who will perform the same task on each object. For example, if `TaskCount` is set to `3` for an image classification labeling job, three workers will classify each input image. Increasing `TaskCount` can improve label accuracy.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 3.

Required: Yes

TaskDescription

A description for the human worker task.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: Yes

TaskTitle

A title for the human worker task.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\t\n\r -\uD7FF\uE000-\uFFFD]*`

Required: Yes

WorkteamArn

Amazon Resource Name (ARN) of a team of workers. To learn more about the types of workforces and work teams you can create and use with Amazon A2I, see Create and Manage Workforces.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:workteam/.*`

Required: Yes

PublicWorkforceTaskPrice

Defines the amount of money paid to an Amazon Mechanical Turk worker for each task performed.

Use one of the following prices for bounding box tasks. Prices are in US dollars and should be based on the complexity of the task; the longer it takes in your initial testing, the more you should offer.

- 0.036
- 0.048
- 0.060

- 0.072
- 0.120
- 0.240
- 0.360
- 0.480
- 0.600
- 0.720
- 0.840
- 0.960
- 1.080
- 1.200

Use one of the following prices for image classification, text classification, and custom tasks. Prices are in US dollars.

- 0.012
- 0.024
- 0.036
- 0.048
- 0.060
- 0.072
- 0.120
- 0.240
- 0.360
- 0.480
- 0.600
- 0.720
- 0.840
- 0.960
- 1.080
- 1.200

Use one of the following prices for semantic segmentation tasks. Prices are in US dollars.

- 0.840
- 0.960
- 1.080
- 1.200

Use one of the following prices for Textract AnalyzeDocument Important Form Key Amazon Augmented AI review tasks. Prices are in US dollars.

- 2.400
- 2.280
- 2.160
- 2.040
- 1.920
- 1.800
- 1.680
- 1.560
- 1.440
- 1.320
- 1.200
- 1.080
- 0.960
- 0.840
- 0.720
- 0.600
- 0.480
- 0.360
- 0.240
- 0.120
- 0.072
- 0.060
- 0.048
- 0.036

- 0.024
- 0.012

Use one of the following prices for Rekognition DetectModerationLabels Amazon Augmented AI review tasks. Prices are in US dollars.

- 1.200
- 1.080
- 0.960
- 0.840
- 0.720
- 0.600
- 0.480
- 0.360
- 0.240
- 0.120
- 0.072
- 0.060
- 0.048
- 0.036
- 0.024
- 0.012

Use one of the following prices for Amazon Augmented AI custom human review tasks. Prices are in US dollars.

- 1.200
- 1.080
- 0.960
- 0.840
- 0.720
- 0.600
- 0.480
- 0.360

- 0.240
- 0.120
- 0.072
- 0.060
- 0.048
- 0.036
- 0.024
- 0.012

Type: PublicWorkforceTaskPrice object

Required: No

TaskAvailabilityLifetimeInSeconds

The length of time that a task remains available for review by human workers.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

TaskKeywords

Keywords used to describe the task so that workers can discover the task.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 1. Maximum length of 30.

Pattern: `[A-Za-z0-9]+( [A-Za-z0-9]+)*`

Required: No

TaskTimeLimitInSeconds

The amount of time that a worker has to complete a task. The default value is 3,600 seconds (1 hour).

Type: Integer

Valid Range: Minimum value of 30.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HumanLoopRequestSource

Service: Amazon SageMaker Service

Container for configuring the source of human task requests.

Contents

AwsManagedHumanLoopRequestSource

Specifies whether Amazon Rekognition or Amazon Textract are used as the integration source. The default field settings and JSON parsing rules are different based on the integration source. Valid values:

Type: String

Valid Values: `AWS/Rekognition/DetectModerationLabels/Image/V3 | AWS/Textract/AnalyzeDocument/Forms/V1`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HumanTaskConfig

Service: Amazon SageMaker Service

Information required for human workers to complete a labeling task.

Contents

NumberOfHumanWorkersPerDataObject

The number of human workers that will label an object.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 9.

Required: Yes

TaskDescription

A description of the task for your human workers.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: Yes

TaskTimeLimitInSeconds

The amount of time that a worker has to complete a task.

If you create a custom labeling job, the maximum value for this parameter is 8 hours (28,800 seconds).

If you create a labeling job using a built-in task type the maximum for this parameter depends on the task type you use:

- For image and text labeling jobs, the maximum is 8 hours (28,800 seconds).
- For 3D point cloud and video frame labeling jobs, the maximum is 30 days (2952,000 seconds) for non-AL mode. For most users, the maximum is also 30 days.

Type: Integer

Valid Range: Minimum value of 30.

Required: Yes

TaskTitle

A title for the task for your human workers.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\t\n\r -\uD7FF\uE000-\uFFFD]*`

Required: Yes

UiConfig

Information about the user interface that workers use to complete the labeling task.

Type: UiConfig object

Required: Yes

WorkteamArn

The Amazon Resource Name (ARN) of the work team assigned to complete the tasks.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:workteam/.*`

Required: Yes

AnnotationConsolidationConfig

Configures how labels are consolidated across human workers.

Type: AnnotationConsolidationConfig object

Required: No

MaxConcurrentTaskCount

Defines the maximum number of data objects that can be labeled by human workers at the same time. Also referred to as batch size. Each object may have more than one worker at

one time. The default value is 1000 objects. To increase the maximum value to 5000 objects, contact AWS Support.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 5000.

Required: No

PreHumanTaskLambdaArn

The Amazon Resource Name (ARN) of a Lambda function that is run before a data object is sent to a human worker. Use this function to provide input to a custom labeling job.

For [built-in task types], use one of the following Amazon SageMaker Ground Truth Lambda function ARNs for `PreHumanTaskLambdaArn`. For custom labeling workflows, see [Pre-annotation Lambda].

Bounding box - Finds the most similar boxes from different workers based on the Jaccard index of the boxes.

- `arn:aws:lambda:us-east-1:432418664414:function:PRE-BoundingBox`
- `arn:aws:lambda:us-east-2:266458841044:function:PRE-BoundingBox`
- `arn:aws:lambda:us-west-2:081040173940:function:PRE-BoundingBox`
- `arn:aws:lambda:ca-central-1:918755190332:function:PRE-BoundingBox`
- `arn:aws:lambda:eu-west-1:568282634449:function:PRE-BoundingBox`
- `arn:aws:lambda:eu-west-2:487402164563:function:PRE-BoundingBox`
- `arn:aws:lambda:eu-central-1:203001061592:function:PRE-BoundingBox`
- `arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-BoundingBox`
- `arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-BoundingBox`
- `arn:aws:lambda:ap-south-1:565803892007:function:PRE-BoundingBox`
- `arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-BoundingBox`
- `arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-BoundingBox`

Image classification - Uses a variant of the Expectation Maximization approach to estimate the true class of an image based on annotations from individual workers.

- `arn:aws:lambda:us-east-1:432418664414:function:PRE-ImageMultiClass`
- `arn:aws:lambda:us-east-2:266458841044:function:PRE-ImageMultiClass`

- arn:aws:lambda:us-west-2:081040173940:function:PRE-ImageMultiClass
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-ImageMultiClass
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-ImageMultiClass
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-ImageMultiClass
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-ImageMultiClass
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-ImageMultiClass
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-ImageMultiClass
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-ImageMultiClass
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-ImageMultiClass
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-ImageMultiClass

Multi-label image classification - Uses a variant of the Expectation Maximization approach to estimate the true classes of an image based on annotations from individual workers.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-ImageMultiClassMultiLabel
- arn:aws:lambda:us-east-2:266458841044:function:PRE-ImageMultiClassMultiLabel
- arn:aws:lambda:us-west-2:081040173940:function:PRE-ImageMultiClassMultiLabel
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-ImageMultiClassMultiLabel
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-ImageMultiClassMultiLabel
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-ImageMultiClassMultiLabel
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-ImageMultiClassMultiLabel
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-ImageMultiClassMultiLabel

- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-ImageMultiClassMultiLabel
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-ImageMultiClassMultiLabel
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-ImageMultiClassMultiLabel
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-ImageMultiClassMultiLabel

Semantic segmentation - Treats each pixel in an image as a multi-class classification and treats pixel annotations from workers as "votes" for the correct label.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-SemanticSegmentation
- arn:aws:lambda:us-east-2:266458841044:function:PRE-SemanticSegmentation
- arn:aws:lambda:us-west-2:081040173940:function:PRE-SemanticSegmentation
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-SemanticSegmentation
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-SemanticSegmentation
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-SemanticSegmentation
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-SemanticSegmentation
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-SemanticSegmentation
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-SemanticSegmentation
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-SemanticSegmentation
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-SemanticSegmentation

- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-SemanticSegmentation

Text classification - Uses a variant of the Expectation Maximization approach to estimate the true class of text based on annotations from individual workers.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-TextMultiClass
- arn:aws:lambda:us-east-2:266458841044:function:PRE-TextMultiClass
- arn:aws:lambda:us-west-2:081040173940:function:PRE-TextMultiClass
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-TextMultiClass
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-TextMultiClass
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-TextMultiClass
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-TextMultiClass
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-TextMultiClass
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-TextMultiClass
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-TextMultiClass
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-TextMultiClass
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-TextMultiClass

Multi-label text classification - Uses a variant of the Expectation Maximization approach to estimate the true classes of text based on annotations from individual workers.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-TextMultiClassMultiLabel
- arn:aws:lambda:us-east-2:266458841044:function:PRE-TextMultiClassMultiLabel
- arn:aws:lambda:us-west-2:081040173940:function:PRE-TextMultiClassMultiLabel
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-TextMultiClassMultiLabel
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-TextMultiClassMultiLabel
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-TextMultiClassMultiLabel

- arn:aws:lambda:eu-central-1:203001061592:function:PRE-TextMultiClassMultiLabel
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-TextMultiClassMultiLabel
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-TextMultiClassMultiLabel
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-TextMultiClassMultiLabel
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-TextMultiClassMultiLabel
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-TextMultiClassMultiLabel

Named entity recognition - Groups similar selections and calculates aggregate boundaries, resolving to most-assigned label.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-NamedEntityRecognition
- arn:aws:lambda:us-east-2:266458841044:function:PRE-NamedEntityRecognition
- arn:aws:lambda:us-west-2:081040173940:function:PRE-NamedEntityRecognition
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-NamedEntityRecognition
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-NamedEntityRecognition
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-NamedEntityRecognition
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-NamedEntityRecognition
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-NamedEntityRecognition
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-NamedEntityRecognition

- arn:aws:lambda:ap-south-1:565803892007:function:PRE-NamedEntityRecognition
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-NamedEntityRecognition
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-NamedEntityRecognition

Video Classification - Use this task type when you need workers to classify videos using predefined labels that you specify. Workers are shown videos and are asked to choose one label for each video.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-VideoMultiClass
- arn:aws:lambda:us-east-2:266458841044:function:PRE-VideoMultiClass
- arn:aws:lambda:us-west-2:081040173940:function:PRE-VideoMultiClass
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-VideoMultiClass
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-VideoMultiClass
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-VideoMultiClass
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-VideoMultiClass
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-VideoMultiClass
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-VideoMultiClass
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-VideoMultiClass
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-VideoMultiClass
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-VideoMultiClass

Video Frame Object Detection - Use this task type to have workers identify and locate objects in a sequence of video frames (images extracted from a video) using bounding boxes. For example, you can use this task to ask workers to identify and localize various objects in a series of video frames, such as cars, bikes, and pedestrians.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-VideoObjectDetection

- arn:aws:lambda:us-east-2:266458841044:function:PRE-VideoObjectDetection
- arn:aws:lambda:us-west-2:081040173940:function:PRE-VideoObjectDetection
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-VideoObjectDetection
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-VideoObjectDetection
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-VideoObjectDetection
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-VideoObjectDetection
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-VideoObjectDetection
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-VideoObjectDetection
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-VideoObjectDetection
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-VideoObjectDetection
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-VideoObjectDetection

Video Frame Object Tracking - Use this task type to have workers track the movement of objects in a sequence of video frames (images extracted from a video) using bounding boxes. For example, you can use this task to ask workers to track the movement of objects, such as cars, bikes, and pedestrians.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-VideoObjectTracking
- arn:aws:lambda:us-east-2:266458841044:function:PRE-VideoObjectTracking
- arn:aws:lambda:us-west-2:081040173940:function:PRE-VideoObjectTracking
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-VideoObjectTracking
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-VideoObjectTracking

- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-VideoObjectTracking
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-VideoObjectTracking
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-VideoObjectTracking
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-VideoObjectTracking
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-VideoObjectTracking
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-VideoObjectTracking
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-VideoObjectTracking

3D Point Cloud Modalities

Use the following pre-annotation lambdas for 3D point cloud labeling modality tasks. See 3D Point Cloud Task types to learn more.

3D Point Cloud Object Detection - Use this task type when you want workers to classify objects in a 3D point cloud by drawing 3D cuboids around objects. For example, you can use this task type to ask workers to identify different types of objects in a point cloud, such as cars, bikes, and pedestrians.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-3DPointCloudObjectDetection
- arn:aws:lambda:us-east-2:266458841044:function:PRE-3DPointCloudObjectDetection
- arn:aws:lambda:us-west-2:081040173940:function:PRE-3DPointCloudObjectDetection
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-3DPointCloudObjectDetection
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-3DPointCloudObjectDetection
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-3DPointCloudObjectDetection

- arn:aws:lambda:ap-south-1:565803892007:function:PRE-3DPointCloudObjectDetection
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-3DPointCloudObjectDetection
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-3DPointCloudObjectDetection
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-3DPointCloudObjectDetection
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-3DPointCloudObjectDetection
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-3DPointCloudObjectDetection

3D Point Cloud Object Tracking - Use this task type when you want workers to draw 3D cuboids around objects that appear in a sequence of 3D point cloud frames. For example, you can use this task type to ask workers to track the movement of vehicles across multiple point cloud frames.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-3DPointCloudObjectTracking
- arn:aws:lambda:us-east-2:266458841044:function:PRE-3DPointCloudObjectTracking
- arn:aws:lambda:us-west-2:081040173940:function:PRE-3DPointCloudObjectTracking
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-3DPointCloudObjectTracking
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-3DPointCloudObjectTracking
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-3DPointCloudObjectTracking
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-3DPointCloudObjectTracking
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-3DPointCloudObjectTracking

- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-3DPointCloudObjectTracking
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-3DPointCloudObjectTracking
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-3DPointCloudObjectTracking
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-3DPointCloudObjectTracking

3D Point Cloud Semantic Segmentation - Use this task type when you want workers to create a point-level semantic segmentation masks by painting objects in a 3D point cloud using different colors where each color is assigned to one of the classes you specify.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-3DPointCloudSemanticSegmentation
- arn:aws:lambda:us-east-2:266458841044:function:PRE-3DPointCloudSemanticSegmentation
- arn:aws:lambda:us-west-2:081040173940:function:PRE-3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-3DPointCloudSemanticSegmentation

- arn:aws:lambda:ca-central-1:918755190332:function:PRE-3DPointCloudSemanticSegmentation

Use the following ARNs for Label Verification and Adjustment Jobs

Use label verification and adjustment jobs to review and adjust labels. To learn more, see Verify and Adjust Labels .

Bounding box verification - Uses a variant of the Expectation Maximization approach to estimate the true class of verification judgement for bounding box labels based on annotations from individual workers.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-VerificationBoundingBox
- arn:aws:lambda:us-east-2:266458841044:function:PRE-VerificationBoundingBox
- arn:aws:lambda:us-west-2:081040173940:function:PRE-VerificationBoundingBox
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-VerificationBoundingBox
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-VerificationBoundingBox
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-VerificationBoundingBox
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-VerificationBoundingBox
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-VerificationBoundingBox
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-VerificationBoundingBox
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-VerificationBoundingBox
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-VerificationBoundingBox
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-VerificationBoundingBox

Bounding box adjustment - Finds the most similar boxes from different workers based on the Jaccard index of the adjusted annotations.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-AdjustmentBoundingBox
- arn:aws:lambda:us-east-2:266458841044:function:PRE-AdjustmentBoundingBox
- arn:aws:lambda:us-west-2:081040173940:function:PRE-AdjustmentBoundingBox
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-AdjustmentBoundingBox
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-AdjustmentBoundingBox
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-AdjustmentBoundingBox
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-AdjustmentBoundingBox
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-AdjustmentBoundingBox
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-AdjustmentBoundingBox
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-AdjustmentBoundingBox
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-AdjustmentBoundingBox
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-AdjustmentBoundingBox

Semantic segmentation verification - Uses a variant of the Expectation Maximization approach to estimate the true class of verification judgment for semantic segmentation labels based on annotations from individual workers.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-VerificationSemanticSegmentation
- arn:aws:lambda:us-east-2:266458841044:function:PRE-VerificationSemanticSegmentation

- arn:aws:lambda:us-west-2:081040173940:function:PRE-VerificationSemanticSegmentation
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-VerificationSemanticSegmentation
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-VerificationSemanticSegmentation
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-VerificationSemanticSegmentation
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-VerificationSemanticSegmentation
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-VerificationSemanticSegmentation
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-VerificationSemanticSegmentation
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-VerificationSemanticSegmentation
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-VerificationSemanticSegmentation
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-VerificationSemanticSegmentation

Semantic segmentation adjustment - Treats each pixel in an image as a multi-class classification and treats pixel adjusted annotations from workers as "votes" for the correct label.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-AdjustmentSemanticSegmentation
- arn:aws:lambda:us-east-2:266458841044:function:PRE-AdjustmentSemanticSegmentation
- arn:aws:lambda:us-west-2:081040173940:function:PRE-AdjustmentSemanticSegmentation
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-AdjustmentSemanticSegmentation
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-AdjustmentSemanticSegmentation

- arn:aws:lambda:eu-west-2:487402164563:function:PRE-AdjustmentSemanticSegmentation
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-AdjustmentSemanticSegmentation
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-AdjustmentSemanticSegmentation
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-AdjustmentSemanticSegmentation
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-AdjustmentSemanticSegmentation
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-AdjustmentSemanticSegmentation
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-AdjustmentSemanticSegmentation

Video Frame Object Detection Adjustment - Use this task type when you want workers to adjust bounding boxes that workers have added to video frames to classify and localize objects in a sequence of video frames.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-AdjustmentVideoObjectDetection
- arn:aws:lambda:us-east-2:266458841044:function:PRE-AdjustmentVideoObjectDetection
- arn:aws:lambda:us-west-2:081040173940:function:PRE-AdjustmentVideoObjectDetection
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-AdjustmentVideoObjectDetection
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-AdjustmentVideoObjectDetection
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-AdjustmentVideoObjectDetection
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-AdjustmentVideoObjectDetection
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-AdjustmentVideoObjectDetection

- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-AdjustmentVideoObjectDetection
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-AdjustmentVideoObjectDetection
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-AdjustmentVideoObjectDetection
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-AdjustmentVideoObjectDetection

Video Frame Object Tracking Adjustment - Use this task type when you want workers to adjust bounding boxes that workers have added to video frames to track object movement across a sequence of video frames.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-AdjustmentVideoObjectTracking
- arn:aws:lambda:us-east-2:266458841044:function:PRE-AdjustmentVideoObjectTracking
- arn:aws:lambda:us-west-2:081040173940:function:PRE-AdjustmentVideoObjectTracking
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-AdjustmentVideoObjectTracking
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-AdjustmentVideoObjectTracking
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-AdjustmentVideoObjectTracking
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-AdjustmentVideoObjectTracking
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-AdjustmentVideoObjectTracking
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-AdjustmentVideoObjectTracking
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-AdjustmentVideoObjectTracking
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-AdjustmentVideoObjectTracking

- arn:aws:lambda:ca-central-1:918755190332:function:PRE-AdjustmentVideoObjectTracking

3D point cloud object detection adjustment - Adjust 3D cuboids in a point cloud frame.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:us-east-2:266458841044:function:PRE-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:us-west-2:081040173940:function:PRE-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-Adjustment3DPointCloudObjectDetection
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-Adjustment3DPointCloudObjectDetection

3D point cloud object tracking adjustment - Adjust 3D cuboids across a sequence of point cloud frames.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-Adjustment3DPointCloudObjectTracking

- arn:aws:lambda:us-east-2:266458841044:function:PRE-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:us-west-2:081040173940:function:PRE-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-Adjustment3DPointCloudObjectTracking
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-Adjustment3DPointCloudObjectTracking

3D point cloud semantic segmentation adjustment - Adjust semantic segmentation masks in a 3D point cloud.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:us-east-2:266458841044:function:PRE-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:us-west-2:081040173940:function:PRE-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-Adjustment3DPointCloudSemanticSegmentation

- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-Adjustment3DPointCloudSemanticSegmentation
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-Adjustment3DPointCloudSemanticSegmentation

Generative AI/Custom - Direct passthrough of input data without any transformation.

- arn:aws:lambda:us-east-1:432418664414:function:PRE-PassThrough
- arn:aws:lambda:us-east-2:266458841044:function:PRE-PassThrough
- arn:aws:lambda:us-west-2:081040173940:function:PRE-PassThrough
- arn:aws:lambda:ca-central-1:918755190332:function:PRE-PassThrough
- arn:aws:lambda:eu-west-1:568282634449:function:PRE-PassThrough
- arn:aws:lambda:eu-west-2:487402164563:function:PRE-PassThrough
- arn:aws:lambda:eu-central-1:203001061592:function:PRE-PassThrough
- arn:aws:lambda:ap-northeast-1:477331159723:function:PRE-PassThrough
- arn:aws:lambda:ap-northeast-2:845288260483:function:PRE-PassThrough
- arn:aws:lambda:ap-south-1:565803892007:function:PRE-PassThrough
- arn:aws:lambda:ap-southeast-1:377565633583:function:PRE-PassThrough
- arn:aws:lambda:ap-southeast-2:454466003867:function:PRE-PassThrough

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:lambda:[a-z0-9\-]*:[0-9]{12}:function:.*`

Required: No

PublicWorkforceTaskPrice

The price that you pay for each task performed by an Amazon Mechanical Turk worker.

Type: PublicWorkforceTaskPrice object

Required: No

TaskAvailabilityLifetimeInSeconds

The length of time that a task remains available for labeling by human workers. The default and maximum values for this parameter depend on the type of workforce you use.

- If you choose the Amazon Mechanical Turk workforce, the maximum is 12 hours (43,200 seconds). The default is 6 hours (21,600 seconds).
- If you choose a private or vendor workforce, the default value is 30 days (2592,000 seconds) for non-AL mode. For most users, the maximum is also 30 days.

Type: Integer

Valid Range: Minimum value of 60.

Required: No

TaskKeywords

Keywords used to describe the task so that workers on Amazon Mechanical Turk can discover the task.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 1. Maximum length of 30.

Pattern: `[A-Za-z0-9]+( [A-Za-z0-9]+)*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HumanTaskUiSummary

Service: Amazon SageMaker Service

Container for human task user interface information.

Contents

CreationTime

A timestamp when SageMaker created the human task user interface.

Type: Timestamp

Required: Yes

HumanTaskUiArn

The Amazon Resource Name (ARN) of the human task user interface.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]+:[0-9]{12}:human-task-ui/.*`

Required: Yes

HumanTaskUiName

The name of the human task user interface.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9])*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperbandStrategyConfig

Service: Amazon SageMaker Service

The configuration for `Hyperband`, a multi-fidelity based hyperparameter tuning strategy. `Hyperband` uses the final and intermediate results of a training job to dynamically allocate resources to utilized hyperparameter configurations while automatically stopping under-performing configurations. This parameter should be provided only if `Hyperband` is selected as the `StrategyConfig` under the `HyperParameterTuningJobConfig` API.

Contents

MaxResource

The maximum number of resources (such as epochs) that can be used by a training job launched by a hyperparameter tuning job. Once a job reaches the `MaxResource` value, it is stopped. If a value for `MaxResource` is not provided, and `Hyperband` is selected as the hyperparameter tuning strategy, `HyperbandTraining` attempts to infer `MaxResource` from the following keys (if present) in StaticsHyperParameters:

- `epochs`
- `numepochs`
- `n-epochs`
- `n_epochs`
- `num_epochs`

If `HyperbandStrategyConfig` is unable to infer a value for `MaxResource`, it generates a validation error. The maximum value is 20,000 epochs. All metrics that correspond to an objective metric are used to derive early stopping decisions. For distributed training jobs, ensure that duplicate metrics are not printed in the logs across the individual nodes in a training job. If multiple nodes are publishing duplicate or incorrect metrics, training jobs may make an incorrect stopping decision and stop the job prematurely.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

MinResource

The minimum number of resources (such as epochs) that can be used by a training job launched by a hyperparameter tuning job. If the value for `MinResource` has not been reached, the training job is not stopped by `Hyperband`.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterAlgorithmSpecification

Service: Amazon SageMaker Service

Specifies which training algorithm to use for training jobs that a hyperparameter tuning job launches and the metrics to monitor.

Contents

TrainingInputMode

The training input mode that the algorithm supports. For more information about input modes, see Algorithms.

Pipe mode

If an algorithm supports `Pipe` mode, Amazon SageMaker streams data directly from Amazon S3 to the container.

File mode

If an algorithm supports `File` mode, SageMaker downloads the training data from S3 to the provisioned ML storage volume, and mounts the directory to the Docker volume for the training container.

You must provision the ML storage volume with sufficient capacity to accommodate the data downloaded from S3. In addition to the training data, the ML storage volume also stores the output model. The algorithm container uses the ML storage volume to also store intermediate information, if any.

For distributed algorithms, training data is distributed uniformly. Your training duration is predictable if the input data objects sizes are approximately the same. SageMaker does not split the files any further for model training. If the object sizes are skewed, training won't be optimal as the data distribution is also skewed when one host in a training cluster is overloaded, thus becoming a bottleneck in training.

FastFile mode

If an algorithm supports `FastFile` mode, SageMaker streams data directly from S3 to the container with no code changes, and provides file system access to the data. Users can author their training script to interact with these files as if they were stored on disk.

`FastFile` mode works best when the data is read sequentially. Augmented manifest files aren't supported. The startup time is lower when there are fewer files in the S3 bucket provided.

Type: String

Valid Values: `Pipe | File | FastFile`

Required: Yes

AlgorithmName

The name of the resource algorithm to use for the hyperparameter tuning job. If you specify a value for this parameter, do not specify a value for `TrainingImage`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 170.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*\/)?([a-zA-Z0-9]([a-zA-Z0-9-]){0,62})(?<!-)`

Required: No

MetricDefinitions

An array of MetricDefinition objects that specify the metrics that the algorithm emits.

Type: Array of MetricDefinition objects

Array Members: Minimum number of 0 items. Maximum number of 40 items.

Required: No

TrainingImage

The registry path of the Docker image that contains the training algorithm. For information about Docker registry paths for built-in algorithms, see Algorithms Provided by Amazon SageMaker: Common Parameters. SageMaker supports both `registry/repository[:tag]` and `registry/repository[@digest]` image path formats. For more information, see Using Your Own Algorithms with Amazon SageMaker.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterSpecification

Service: Amazon SageMaker Service

Defines a hyperparameter to be used by an algorithm.

Contents

Name

The name of this hyperparameter. The name must be unique.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: Yes

Type

The type of this hyperparameter. The valid types are `Integer`, `Continuous`, `Categorical`, and `FreeText`.

Type: String

Valid Values: `Integer | Continuous | Categorical | FreeText`

Required: Yes

DefaultValue

The default value for this hyperparameter. If a default value is specified, a hyperparameter cannot be required.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2500.

Pattern: `.*`

Required: No

Description

A brief description of the hyperparameter.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

IsRequired

Indicates whether this hyperparameter is required.

Type: Boolean

Required: No

IsTunable

Indicates whether this hyperparameter is tunable in a hyperparameter tuning job.

Type: Boolean

Required: No

Range

The allowed range for this hyperparameter.

Type: ParameterRange object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTrainingJobDefinition

Service: Amazon SageMaker Service

Defines the training jobs launched by a hyperparameter tuning job.

Contents

AlgorithmSpecification

The HyperParameterAlgorithmSpecification object that specifies the resource algorithm to use for the training jobs that the tuning job launches.

Type: HyperParameterAlgorithmSpecification object

Required: Yes

OutputDataConfig

Specifies the path to the Amazon S3 bucket where you store model artifacts from the training jobs that the tuning job launches.

Type: OutputDataConfig object

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of the IAM role associated with the training jobs that the tuning job launches.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

StoppingCondition

Specifies a limit to how long a model hyperparameter training job can run. It also specifies how long a managed spot training job has to complete. When the job reaches the time limit, SageMaker ends the training job. Use this API to cap model training costs.

Type: StoppingCondition object

Required: Yes

CheckpointConfig

Contains information about the output location for managed spot training checkpoint data.

Type: CheckpointConfig object

Required: No

DefinitionName

The job definition name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: No

EnableInterContainerTrafficEncryption

To encrypt all communications between ML compute instances in distributed training, choose `True`. Encryption provides greater security for distributed training, but training might take longer. How long it takes depends on the amount of communication between compute instances, especially if you use a deep learning algorithm in distributed training.

Type: Boolean

Required: No

EnableManagedSpotTraining

A Boolean indicating whether managed spot training is enabled (`True`) or not (`False`).

Type: Boolean

Required: No

EnableNetworkIsolation

Isolates the training container. No inbound or outbound network calls can be made, except for calls between peers within a training cluster for distributed training. If network isolation is used for training jobs that are configured to use a VPC, SageMaker downloads and uploads customer

data and model artifacts through the specified VPC, but the training container does not have network access.

Type: Boolean

Required: No

Environment

An environment variable that you can pass into the SageMaker CreateTrainingJob API. You can use an existing environment variable from the training container or use your own. See Define metrics and variables for more information.

> **Note**
>
> The maximum number of items specified for `Map Entries` refers to the maximum number of environment variables for each `TrainingJobDefinition` and also the maximum for the hyperparameter tuning job itself. That is, the sum of the number of environment variables for all the training job definitions can't exceed the maximum number specified.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 48 items.

Key Length Constraints: Minimum length of 0. Maximum length of 512.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 512.

Value Pattern: `[\S\s]*`

Required: No

HyperParameterRanges

Specifies ranges of integer, continuous, and categorical hyperparameters that a hyperparameter tuning job searches. The hyperparameter tuning job launches training jobs with hyperparameter values within these ranges to find the combination of values that result in the training job with the best performance as measured by the objective metric of the hyperparameter tuning job.

> **Note**
>
> The maximum number of items specified for `Array Members` refers to the maximum number of hyperparameters for each range and also the maximum for the hyperparameter tuning job itself. That is, the sum of the number of hyperparameters for all the ranges can't exceed the maximum number specified.

Type: ParameterRanges object

Required: No

HyperParameterTuningResourceConfig

The configuration for the hyperparameter tuning resources, including the compute instances and storage volumes, used for training jobs launched by the tuning job. By default, storage volumes hold model artifacts and incremental states. Choose `File` for `TrainingInputMode` in the `AlgorithmSpecification` parameter to additionally store training data in the storage volume (optional).

Type: HyperParameterTuningResourceConfig object

Required: No

InputDataConfig

An array of Channel objects that specify the input for the training jobs that the tuning job launches.

Type: Array of Channel objects

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Required: No

ResourceConfig

The resources, including the compute instances and storage volumes, to use for the training jobs that the tuning job launches.

Storage volumes store model artifacts and incremental states. Training algorithms might also use storage volumes for scratch space. If you want SageMaker to use the storage volume

to store the training data, choose `File` as the `TrainingInputMode` in the algorithm specification. For distributed training algorithms, specify an instance count greater than 1.

> **Note**
>
> If you want to use hyperparameter optimization with instance type flexibility, use `HyperParameterTuningResourceConfig` instead.

Type: ResourceConfig object

Required: No

RetryStrategy

The number of times to retry the job when the job fails due to an `InternalServerError`.

Type: RetryStrategy object

Required: No

StaticHyperParameters

Specifies the values of hyperparameters that do not change for the tuning job.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Required: No

TuningObjective

Defines the objective metric for a hyperparameter tuning job. Hyperparameter tuning uses the value of this metric to evaluate the training jobs it launches, and returns the training job that results in either the highest or lowest value for this metric, depending on the value you specify

for the Type parameter. If you want to define a custom objective metric, see Define metrics and environment variables.

Type: HyperParameterTuningJobObjective object

Required: No

VpcConfig

The VpcConfig object that specifies the VPC that you want the training jobs that this hyperparameter tuning job launches to connect to. Control access to and from your training container by configuring the VPC. For more information, see Protect Training Jobs by Using an Amazon Virtual Private Cloud.

Type: VpcConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTrainingJobSummary

Service: Amazon SageMaker Service

The container for the summary information about a training job.

Contents

CreationTime

The date and time that the training job was created.

Type: Timestamp

Required: Yes

TrainingJobArn

The Amazon Resource Name (ARN) of the training job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

TrainingJobName

The name of the training job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

TrainingJobStatus

The status of the training job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting`

Required: Yes

TunedHyperParameters

A list of the hyperparameters for which you specified ranges to search.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Required: Yes

FailureReason

The reason that the training job failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

FinalHyperParameterTuningJobObjectiveMetric

The FinalHyperParameterTuningJobObjectiveMetric object that specifies the value of the objective metric of the tuning job that launched this training job.

Type: FinalHyperParameterTuningJobObjectiveMetric object

Required: No

ObjectiveStatus

The status of the objective metric for the training job:

- Succeeded: The final objective metric for the training job was evaluated by the hyperparameter tuning job and used in the hyperparameter tuning process.
- Pending: The training job is in progress and evaluation of its final objective metric is pending.
- Failed: The final objective metric for the training job was not evaluated, and was not used in the hyperparameter tuning process. This typically occurs when the training job failed or did not emit an objective metric.

Type: String

Valid Values: `Succeeded | Pending | Failed`

Required: No

TrainingEndTime

Specifies the time when the training job ends on training instances. You are billed for the time interval between the value of `TrainingStartTime` and this time. For successful jobs and stopped jobs, this is the time after model artifacts are uploaded. For failed jobs, this is the time when SageMaker detects a job failure.

Type: Timestamp

Required: No

TrainingJobDefinitionName

The training job definition name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: No

TrainingStartTime

The date and time that the training job started.

Type: Timestamp

Required: No

TuningJobName

The HyperParameter tuning job that launched the training job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTuningInstanceConfig

Service: Amazon SageMaker Service

The configuration for hyperparameter tuning resources for use in training jobs launched by the tuning job. These resources include compute instances and storage volumes. Specify one or more compute instance configurations and allocation strategies to select resources (optional).

Contents

InstanceCount

The number of instances of the type specified by `InstanceType`. Choose an instance count larger than 1 for distributed training algorithms. See Step 2: Launch a SageMaker Distributed Training Job Using the SageMaker Python SDK for more information.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

InstanceType

The instance type used for processing of hyperparameter optimization jobs. Choose from general purpose (no GPUs) instance types: ml.m5.xlarge, ml.m5.2xlarge, and ml.m5.4xlarge or compute optimized (no GPUs) instance types: ml.c5.xlarge and ml.c5.2xlarge. For more information about instance types, see instance type descriptions.

Type: String

Valid Values: `ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p3dn.24xlarge | ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5n.xlarge | ml.c5n.2xlarge`

| ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.8xlarge | ml.c6i.4xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.8xlarge | ml.r5d.12xlarge | ml.r5d.16xlarge | ml.r5d.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.p6-b200.48xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge | ml.p6-b300.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge

Required: Yes

VolumeSizeInGB

The volume size in GB of the data to be processed for hyperparameter optimization (optional).

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTuningJobCompletionDetails

Service: Amazon SageMaker Service

A structure that contains runtime information about both current and completed hyperparameter tuning jobs.

Contents

ConvergenceDetectedTime

The time in timestamp format that AMT detected model convergence, as defined by a lack of significant improvement over time based on criteria developed over a wide range of diverse benchmarking tests.

Type: Timestamp

Required: No

NumberOfTrainingJobsObjectiveNotImproving

The number of training jobs launched by a tuning job that are not improving (1% or less) as measured by model performance evaluated against an objective function.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTuningJobConfig

Service: Amazon SageMaker Service

Configures a hyperparameter tuning job.

Contents

ResourceLimits

The ResourceLimits object that specifies the maximum number of training and parallel training jobs that can be used for this hyperparameter tuning job.

Type: ResourceLimits object

Required: Yes

Strategy

Specifies how hyperparameter tuning chooses the combinations of hyperparameter values to use for the training job it launches. For information about search strategies, see How Hyperparameter Tuning Works.

Type: String

Valid Values: `Bayesian | Random | Hyperband | Grid`

Required: Yes

HyperParameterTuningJobObjective

The HyperParameterTuningJobObjective specifies the objective metric used to evaluate the performance of training jobs launched by this tuning job.

Type: HyperParameterTuningJobObjective object

Required: No

ParameterRanges

The ParameterRanges object that specifies the ranges of hyperparameters that this tuning job searches over to find the optimal configuration for the highest model performance against your chosen objective metric.

Type: ParameterRanges object

Required: No

RandomSeed

A value used to initialize a pseudo-random number generator. Setting a random seed and using the same seed later for the same tuning job will allow hyperparameter optimization to find more a consistent hyperparameter configuration between the two runs.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

StrategyConfig

The configuration for the `Hyperband` optimization strategy. This parameter should be provided only if `Hyperband` is selected as the strategy for `HyperParameterTuningJobConfig`.

Type: HyperParameterTuningJobStrategyConfig object

Required: No

TrainingJobEarlyStoppingType

Specifies whether to use early stopping for training jobs launched by the hyperparameter tuning job. Because the `Hyperband` strategy has its own advanced internal early stopping mechanism, `TrainingJobEarlyStoppingType` must be `OFF` to use `Hyperband`. This parameter can take on one of the following values (the default value is `OFF`):

OFF

Training jobs launched by the hyperparameter tuning job do not use early stopping.

AUTO

SageMaker stops training jobs launched by the hyperparameter tuning job when they are unlikely to perform better than previously completed training jobs. For more information, see Stop Training Jobs Early.

Type: String

Valid Values: `Off | Auto`

Required: No

TuningJobCompletionCriteria

The tuning job's completion criteria.

Type: TuningJobCompletionCriteria object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTuningJobConsumedResources

Service: Amazon SageMaker Service

The total resources consumed by your hyperparameter tuning job.

Contents

RuntimeInSeconds

The wall clock runtime in seconds used by your hyperparameter tuning job.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTuningJobObjective

Service: Amazon SageMaker Service

Defines the objective metric for a hyperparameter tuning job. Hyperparameter tuning uses the value of this metric to evaluate the training jobs it launches, and returns the training job that results in either the highest or lowest value for this metric, depending on the value you specify for the Type parameter. If you want to define a custom objective metric, see Define metrics and environment variables.

Contents

MetricName

The name of the metric to use for the objective metric.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: Yes

Type

Whether to minimize or maximize the objective metric.

Type: String

Valid Values: `Maximize | Minimize`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTuningJobSearchEntity

Service: Amazon SageMaker Service

An entity returned by the SearchRecord API containing the properties of a hyperparameter tuning job.

Contents

BestTrainingJob

The container for the summary information about a training job.

Type: HyperParameterTrainingJobSummary object

Required: No

ConsumedResources

The total amount of resources consumed by a hyperparameter tuning job.

Type: HyperParameterTuningJobConsumedResources object

Required: No

CreationTime

The time that a hyperparameter tuning job was created.

Type: Timestamp

Required: No

FailureReason

The error that was created when a hyperparameter tuning job failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

HyperParameterTuningEndTime

The time that a hyperparameter tuning job ended.

Type: Timestamp

Required: No

HyperParameterTuningJobArn

The Amazon Resource Name (ARN) of a hyperparameter tuning job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:hyper-parameter-tuning-job/.*`

Required: No

HyperParameterTuningJobConfig

Configures a hyperparameter tuning job.

Type: HyperParameterTuningJobConfig object

Required: No

HyperParameterTuningJobName

The name of a hyperparameter tuning job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: No

HyperParameterTuningJobStatus

The status of a hyperparameter tuning job.

Type: String

Valid Values: `Completed | InProgress | Failed | Stopped | Stopping | Deleting | DeleteFailed`

Required: No

LastModifiedTime

The time that a hyperparameter tuning job was last modified.

Type: Timestamp

Required: No

ObjectiveStatusCounters

Specifies the number of training jobs that this hyperparameter tuning job launched, categorized by the status of their objective metric. The objective metric status shows whether the final objective metric for the training job has been evaluated by the tuning job and used in the hyperparameter tuning process.

Type: ObjectiveStatusCounters object

Required: No

OverallBestTrainingJob

The container for the summary information about a training job.

Type: HyperParameterTrainingJobSummary object

Required: No

Tags

The tags associated with a hyperparameter tuning job. For more information see Tagging AWS resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TrainingJobDefinition

Defines the training jobs launched by a hyperparameter tuning job.

Type: HyperParameterTrainingJobDefinition object

Required: No

TrainingJobDefinitions

The job definitions included in a hyperparameter tuning job.

Type: Array of HyperParameterTrainingJobDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

TrainingJobStatusCounters

The numbers of training jobs launched by a hyperparameter tuning job, categorized by status.

Type: TrainingJobStatusCounters object

Required: No

TuningJobCompletionDetails

Information about either a current or completed hyperparameter tuning job.

Type: HyperParameterTuningJobCompletionDetails object

Required: No

WarmStartConfig

Specifies the configuration for a hyperparameter tuning job that uses one or more previous hyperparameter tuning jobs as a starting point. The results of previous tuning jobs are used to inform which combinations of hyperparameters to search over in the new tuning job.

All training jobs launched by the new hyperparameter tuning job are evaluated by using the objective metric, and the training job that performs the best is compared to the best training jobs from the parent tuning jobs. From these, the training job that performs the best as measured by the objective metric is returned as the overall best training job.

> **Note**
>
> All training jobs launched by parent hyperparameter tuning jobs and the new hyperparameter tuning jobs count against the limit of training jobs for the tuning job.

Type: HyperParameterTuningJobWarmStartConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTuningJobStrategyConfig

Service: Amazon SageMaker Service

The configuration for a training job launched by a hyperparameter tuning job. Choose `Bayesian` for Bayesian optimization, and `Random` for random search optimization. For more advanced use cases, use `Hyperband`, which evaluates objective metrics for training jobs after every epoch. For more information about strategies, see How Hyperparameter Tuning Works.

Contents

HyperbandStrategyConfig

The configuration for the object that specifies the `Hyperband` strategy. This parameter is only supported for the `Hyperband` selection for `Strategy` within the `HyperParameterTuningJobConfig` API.

Type: HyperbandStrategyConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTuningJobSummary

Service: Amazon SageMaker Service

Provides summary information about a hyperparameter tuning job.

Contents

CreationTime

The date and time that the tuning job was created.

Type: Timestamp

Required: Yes

HyperParameterTuningJobArn

The Amazon Resource Name (ARN) of the tuning job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:hyper-parameter-tuning-job/.*`

Required: Yes

HyperParameterTuningJobName

The name of the tuning job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

HyperParameterTuningJobStatus

The status of the tuning job.

Type: String

Valid Values: `Completed | InProgress | Failed | Stopped | Stopping | Deleting | DeleteFailed`

Required: Yes

ObjectiveStatusCounters

The ObjectiveStatusCounters object that specifies the numbers of training jobs, categorized by objective metric status, that this tuning job launched.

Type: ObjectiveStatusCounters object

Required: Yes

Strategy

Specifies the search strategy hyperparameter tuning uses to choose which hyperparameters to evaluate at each iteration.

Type: String

Valid Values: `Bayesian | Random | Hyperband | Grid`

Required: Yes

TrainingJobStatusCounters

The TrainingJobStatusCounters object that specifies the numbers of training jobs, categorized by status, that this tuning job launched.

Type: TrainingJobStatusCounters object

Required: Yes

HyperParameterTuningEndTime

The date and time that the tuning job ended.

Type: Timestamp

Required: No

LastModifiedTime

The date and time that the tuning job was modified.

Type: Timestamp

Required: No

ResourceLimits

The ResourceLimits object that specifies the maximum number of training jobs and parallel training jobs allowed for this tuning job.

Type: ResourceLimits object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTuningJobWarmStartConfig

Service: Amazon SageMaker Service

Specifies the configuration for a hyperparameter tuning job that uses one or more previous hyperparameter tuning jobs as a starting point. The results of previous tuning jobs are used to inform which combinations of hyperparameters to search over in the new tuning job.

All training jobs launched by the new hyperparameter tuning job are evaluated by using the objective metric, and the training job that performs the best is compared to the best training jobs from the parent tuning jobs. From these, the training job that performs the best as measured by the objective metric is returned as the overall best training job.

> **Note**
>
> All training jobs launched by parent hyperparameter tuning jobs and the new hyperparameter tuning jobs count against the limit of training jobs for the tuning job.

Contents

ParentHyperParameterTuningJobs

An array of hyperparameter tuning jobs that are used as the starting point for the new hyperparameter tuning job. For more information about warm starting a hyperparameter tuning job, see Using a Previous Hyperparameter Tuning Job as a Starting Point.

Hyperparameter tuning jobs created before October 1, 2018 cannot be used as parent jobs for warm start tuning jobs.

Type: Array of ParentHyperParameterTuningJob objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

WarmStartType

Specifies one of the following:

IDENTICAL_DATA_AND_ALGORITHM

The new hyperparameter tuning job uses the same input data and training image as the parent tuning jobs. You can change the hyperparameter ranges to search and the

maximum number of training jobs that the hyperparameter tuning job launches. You cannot use a new version of the training algorithm, unless the changes in the new version do not affect the algorithm itself. For example, changes that improve logging or adding support for a different data format are allowed. You can also change hyperparameters from tunable to static, and from static to tunable, but the total number of static plus tunable hyperparameters must remain the same as it is in all parent jobs. The objective metric for the new tuning job must be the same as for all parent jobs.

TRANSFER_LEARNING

The new hyperparameter tuning job can include input data, hyperparameter ranges, maximum number of concurrent training jobs, and maximum number of training jobs that are different than those of its parent hyperparameter tuning jobs. The training image can also be a different version from the version used in the parent hyperparameter tuning job. You can also change hyperparameters from tunable to static, and from static to tunable, but the total number of static plus tunable hyperparameters must remain the same as it is in all parent jobs. The objective metric for the new tuning job must be the same as for all parent jobs.

Type: String

Valid Values: `IdenticalDataAndAlgorithm | TransferLearning`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HyperParameterTuningResourceConfig

Service: Amazon SageMaker Service

The configuration of resources, including compute instances and storage volumes for use in training jobs launched by hyperparameter tuning jobs. `HyperParameterTuningResourceConfig` is similar to `ResourceConfig`, but has the additional `InstanceConfigs` and `AllocationStrategy` fields to allow for flexible instance management. Specify one or more instance types, count, and the allocation strategy for instance selection.

> **Note**
>
> `HyperParameterTuningResourceConfig` supports the capabilities of `ResourceConfig` with the exception of `KeepAlivePeriodInSeconds`. Hyperparameter tuning jobs use warm pools by default, which reuse clusters between training jobs.

Contents

AllocationStrategy

The strategy that determines the order of preference for resources specified in `InstanceConfigs` used in hyperparameter optimization.

Type: String

Valid Values: `Prioritized`

Required: No

InstanceConfigs

A list containing the configuration(s) for one or more resources for processing hyperparameter jobs. These resources include compute instances and storage volumes to use in model training jobs launched by hyperparameter tuning jobs. The `AllocationStrategy` controls the order in which multiple configurations provided in `InstanceConfigs` are used.

> **Note**
>
> If you only want to use a single instance configuration inside the `HyperParameterTuningResourceConfig` API, do not provide a value for

`InstanceConfigs`. Instead, use `InstanceType`, `VolumeSizeInGB` and `InstanceCount`. If you use `InstanceConfigs`, do not provide values for `InstanceType`, `VolumeSizeInGB` or `InstanceCount`.

Type: Array of HyperParameterTuningInstanceConfig objects

Array Members: Minimum number of 1 item. Maximum number of 6 items.

Required: No

InstanceCount

The number of compute instances of type `InstanceType` to use. For distributed training, select a value greater than 1.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

InstanceType

The instance type used to run hyperparameter optimization tuning jobs. See descriptions of instance types for more information.

Type: String

Valid Values: `ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p3dn.24xlarge | ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5n.xlarge | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge`

```
| ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge
| ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge |
ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge |
ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge |
ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.trn1.2xlarge |
ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.m6i.large
| ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge |
ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge
| ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.8xlarge | ml.c6i.4xlarge |
ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge
| ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge |
ml.r5d.8xlarge | ml.r5d.12xlarge | ml.r5d.16xlarge | ml.r5d.24xlarge |
ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.r5.large
| ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge |
ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.p6-b200.48xlarge
| ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge |
ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge
| ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge |
ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge
| ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge
| ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge
| ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.p6e-
gb200.36xlarge | ml.p5.4xlarge | ml.p6-b300.48xlarge | ml.g7e.2xlarge
| ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge |
ml.g7e.48xlarge
```

Required: No

VolumeKmsKeyId

A key used by AWS Key Management Service to encrypt data on the storage volume attached to the compute instances used to run the training job. You can use either of the following formats to specify a key.

KMS Key ID:

```
"1234abcd-12ab-34cd-56ef-1234567890ab"
```

Amazon Resource Name (ARN) of a KMS key:

`"arn:aws:kms:us-west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab"`

Some instances use local storage, which use a hardware module to encrypt storage volumes. If you choose one of these instance types, you cannot request a `VolumeKmsKeyId`. For a list of instance types that use local storage, see instance store volumes. For more information about AWS Key Management Service, see KMS encryption for more information.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

VolumeSizeInGB

The volume size in GB for the storage volume to be used in processing hyperparameter optimization jobs (optional). These volumes store model artifacts, incremental states and optionally, scratch space for training algorithms. Do not provide a value for this parameter if a value for `InstanceConfigs` is also specified.

Some instance types have a fixed total local storage size. If you select one of these instances for training, `VolumeSizeInGB` cannot be greater than this total size. For a list of instance types with local instance storage and their sizes, see instance store volumes.

> **Note**
>
> SageMaker supports only the General Purpose SSD (gp2) storage volume type.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IamIdentity

Service: Amazon SageMaker Service

The IAM Identity details associated with the user. These details are associated with model package groups, model packages and project entities only.

Contents

Arn

The Amazon Resource Name (ARN) of the IAM identity.

Type: String

Required: No

PrincipalId

The ID of the principal that assumes the IAM identity.

Type: String

Required: No

SourceIdentity

The person or application which assumes the IAM identity.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IamPolicyConstraints

Service: Amazon SageMaker Service

Use this parameter to specify a supported global condition key that is added to the IAM policy.

Contents

SourceIp

When `SourceIp` is `Enabled` the worker's IP address when a task is rendered in the worker portal is added to the IAM policy as a `Condition` used to generate the Amazon S3 presigned URL. This IP address is checked by Amazon S3 and must match in order for the Amazon S3 resource to be rendered in the worker portal.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

VpcSourceIp

When `VpcSourceIp` is `Enabled` the worker's IP address when a task is rendered in private worker portal inside the VPC is added to the IAM policy as a `Condition` used to generate the Amazon S3 presigned URL. To render the task successfully Amazon S3 checks that the presigned URL is being accessed over an Amazon S3 VPC Endpoint, and that the worker's IP address matches the IP address in the IAM policy. To learn more about configuring private worker portal, see Use Amazon VPC mode from a private worker portal.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

IdentityProviderOAuthSetting

Service: Amazon SageMaker Service

The Amazon SageMaker Canvas application setting where you configure OAuth for connecting to an external data source, such as Snowflake.

Contents

DataSourceName

The name of the data source that you're connecting to. Canvas currently supports OAuth for Snowflake and Salesforce Data Cloud.

Type: String

Valid Values: `SalesforceGenie | Snowflake`

Required: No

SecretArn

The ARN of an AWS Secrets Manager secret that stores the credentials from your identity provider, such as the client ID and secret, authorization URL, and token URL.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:secretsmanager:[a-z0-9\-]*:[0-9]{12}:secret:.*`

Required: No

Status

Describes whether OAuth for a data source is enabled or disabled in the Canvas application.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IdleSettings

Service: Amazon SageMaker Service

Settings related to idle shutdown of Studio applications.

Contents

IdleTimeoutInMinutes

The time that SageMaker waits after the application becomes idle before shutting it down.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 525600.

Required: No

LifecycleManagement

Indicates whether idle shutdown is activated for the application type.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

MaxIdleTimeoutInMinutes

The maximum value in minutes that custom idle shutdown can be set to by the user.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 525600.

Required: No

MinIdleTimeoutInMinutes

The minimum value in minutes that custom idle shutdown can be set to by the user.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 525600.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Image

Service: Amazon SageMaker Service

A SageMaker AI image. A SageMaker AI image represents a set of container images that are derived from a common base container image. Each of these container images is represented by a SageMaker AI `ImageVersion`.

Contents

CreationTime

When the image was created.

Type: Timestamp

Required: Yes

ImageArn

The ARN of the image.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image/[a-zA-Z0-9]([-.]?[a-zA-Z0-9])*`

Required: Yes

ImageName

The name of the image.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([-.]?[a-zA-Z0-9]){0,62}`

Required: Yes

ImageStatus

The status of the image.

Type: String

Valid Values: `CREATING | CREATED | CREATE_FAILED | UPDATING | UPDATE_FAILED | DELETING | DELETE_FAILED`

Required: Yes

LastModifiedTime

When the image was last modified.

Type: Timestamp

Required: Yes

Description

The description of the image.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 512.

Pattern: `.*`

Required: No

DisplayName

The name of the image as displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `\S(.*\S)?`

Required: No

FailureReason

When a create, update, or delete operation fails, the reason for the failure.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageClassificationJobConfig

Service: Amazon SageMaker Service

The collection of settings used by an AutoML job V2 for the image classification problem type.

Contents

CompletionCriteria

How long a job is allowed to run, or how many candidates a job is allowed to generate.

Type: AutoMLJobCompletionCriteria object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageConfig

Service: Amazon SageMaker Service

Specifies whether the model container is in Amazon ECR or a private Docker registry accessible from your Amazon Virtual Private Cloud (VPC).

Contents

RepositoryAccessMode

Set this to one of the following values:

- `Platform` - The model image is hosted in Amazon ECR.
- `Vpc` - The model image is hosted in a private Docker registry in your VPC.

Type: String

Valid Values: `Platform | Vpc`

Required: Yes

RepositoryAuthConfig

(Optional) Specifies an authentication configuration for the private docker registry where your model image is hosted. Specify a value for this property only if you specified `Vpc` as the value for the `RepositoryAccessMode` field, and the private Docker registry where the model image is hosted requires authentication.

Type: RepositoryAuthConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageVersion

Service: Amazon SageMaker Service

A version of a SageMaker AI `Image`. A version represents an existing container image.

Contents

CreationTime

When the version was created.

Type: Timestamp

Required: Yes

ImageArn

The ARN of the image the version is based on.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image/[a-zA-Z0-9]([-.]?[a-zA-Z0-9])*`

Required: Yes

ImageVersionArn

The ARN of the version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image-version/[a-z0-9]([-.]?[a-z0-9])*/[0-9]+|None)`

Required: Yes

ImageVersionStatus

The status of the version.

Type: String

Valid Values: `CREATING | CREATED | CREATE_FAILED | DELETING | DELETE_FAILED`

Required: Yes

LastModifiedTime

When the version was last modified.

Type: Timestamp

Required: Yes

Version

The version number.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

FailureReason

When a create or delete operation fails, the reason for the failure.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentAvailabilityZoneBalance

Service: Amazon SageMaker Service

Configuration for balancing inference component copies across Availability Zones.

Contents

EnforcementMode

Determines how strictly the Availability Zone balance constraint is enforced.

PERMISSIVE

The endpoint attempts to balance copies across Availability Zones but proceeds with scheduling even if balance can't be achieved due to available capacity or instance distribution across Availability Zones.

Type: String

Valid Values: `PERMISSIVE`

Required: Yes

MaxImbalance

The maximum allowed difference in the number of inference component copies between any two Availability Zones. This parameter applies only when the endpoint has instances across two or more Availability Zones. A copy placement is allowed if it reduces imbalance or the resulting imbalance is within this value.

Default value: `0`.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentCapacitySize

Service: Amazon SageMaker Service

Specifies the type and size of the endpoint capacity to activate for a rolling deployment or a rollback strategy. You can specify your batches as either of the following:

- A count of inference component copies
- The overall percentage or your fleet

For a rollback strategy, if you don't specify the fields in this object, or if you set the `Value` parameter to 100%, then SageMaker AI uses a blue/green rollback strategy and rolls all traffic back to the blue fleet.

Contents

Type

Specifies the endpoint capacity type.

COPY_COUNT

The endpoint activates based on the number of inference component copies.

CAPACITY_PERCENT

The endpoint activates based on the specified percentage of capacity.

Type: String

Valid Values: `COPY_COUNT | CAPACITY_PERCENT`

Required: Yes

Value

Defines the capacity size, either as a number of inference component copies or a capacity percentage.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentComputeResourceRequirements

Service: Amazon SageMaker Service

Defines the compute resources to allocate to run a model, plus any adapter models, that you assign to an inference component. These resources include CPU cores, accelerators, and memory.

Contents

MinMemoryRequiredInMb

The minimum MB of memory to allocate to run a model that you assign to an inference component.

Type: Integer

Valid Range: Minimum value of 128.

Required: Yes

MaxMemoryRequiredInMb

The maximum MB of memory to allocate to run a model that you assign to an inference component.

Type: Integer

Valid Range: Minimum value of 128.

Required: No

NumberOfAcceleratorDevicesRequired

The number of accelerators to allocate to run a model that you assign to an inference component. Accelerators include GPUs and AWS Inferentia.

Type: Float

Valid Range: Minimum value of 1.

Required: No

NumberOfCpuCoresRequired

The number of CPU cores to allocate to run a model that you assign to an inference component.

Type: Float

Valid Range: Minimum value of 0.25.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentContainerSpecification

Service: Amazon SageMaker Service

Defines a container that provides the runtime environment for a model that you deploy with an inference component.

Contents

ArtifactUrl

The Amazon S3 path where the model artifacts, which result from model training, are stored. This path must point to a single gzip compressed tar archive (.tar.gz suffix).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

ContainerMetricsConfig

The configuration for container metrics scraping. Specifies the metrics endpoint path and publishing frequency for the inference component's container. If not specified when `EnableDetailedObservability` is `True`, the default path `/metrics` on port `8080` is used. For first-party and Deep Learning Containers (DLC), the endpoint path is determined automatically and this configuration is optional.

Type: ContainerMetricsConfig object

Required: No

Environment

The environment variables to set in the Docker container. Each key and value in the Environment string-to-string map can have length of up to 1024. We support up to 16 entries in the map.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 1024.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 1024.

Value Pattern: `[\S\s]*`

Required: No

Image

The Amazon Elastic Container Registry (Amazon ECR) path where the Docker image for the model is stored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[\S]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentContainerSpecificationSummary

Service: Amazon SageMaker Service

Details about the resources that are deployed with this inference component.

Contents

ArtifactUrl

The Amazon S3 path where the model artifacts are stored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

ContainerMetricsConfig

The container metrics scraping configuration for this inference component, including the metrics endpoint path and publishing frequency.

Type: ContainerMetricsConfig object

Required: No

DeployedImage

Gets the Amazon EC2 Container Registry path of the docker image of the model that is hosted in this ProductionVariant.

If you used the `registry/repository[:tag]` form to specify the image path of the primary container when you created the model hosted in this `ProductionVariant`, the path resolves to a path of the form `registry/repository[@digest]`. A digest is a hash value that identifies a specific version of an image. For information about Amazon ECR paths, see Pulling an Image in the *Amazon ECR User Guide*.

Type: DeployedImage object

Required: No

Environment

The environment variables to set in the Docker container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 1024.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 1024.

Value Pattern: `[\S\s]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentDataCacheConfig

Service: Amazon SageMaker Service

Settings that affect how the inference component caches data.

Contents

EnableCaching

Sets whether the endpoint that hosts the inference component caches the model artifacts and container image.

With caching enabled, the endpoint caches this data in each instance that it provisions for the inference component. That way, the inference component deploys faster during the auto scaling process. If caching isn't enabled, the inference component takes longer to deploy because of the time it spends downloading the data.

Type: Boolean

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentDataCacheConfigSummary

Service: Amazon SageMaker Service

Settings that affect how the inference component caches data.

Contents

EnableCaching

Indicates whether the inference component caches model artifacts as part of the auto scaling process.

Type: Boolean

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentDeploymentConfig

Service: Amazon SageMaker Service

The deployment configuration for an endpoint that hosts inference components. The configuration includes the desired deployment strategy and rollback settings.

Contents

RollingUpdatePolicy

Specifies a rolling deployment strategy for updating a SageMaker AI endpoint.

Type: InferenceComponentRollingUpdatePolicy object

Required: Yes

AutoRollbackConfiguration

Automatic rollback configuration for handling endpoint deployment failures and recovery.

Type: AutoRollbackConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentMetadata

Service: Amazon SageMaker Service

The metadata of the inference component.

Contents

Arn

The Amazon Resource Name (ARN) of the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentPlacementStatus

Service: Amazon SageMaker Service

The placement status of an inference component on a specific instance type. Shows the number of inference component copies currently placed on instances of a given type.

Contents

CurrentCopyCount

The number of inference component copies currently placed on instances of this type.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

InstanceType

The ML compute instance type where the inference component copies are placed.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large`

| ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large | ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge | ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large | ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge | ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge | ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge | ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge | ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge | ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge | ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large | ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge | ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge | ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge | ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge | ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge

```
| ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge |
ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge
| ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge |
ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge |
ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium
| ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge |
ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge
| ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge |
ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge
| ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge |
ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge |
ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large |
ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge
| ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge |
ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-
gb200.36xlarge | ml.p5.4xlarge
```

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentRollingUpdatePolicy

Service: Amazon SageMaker Service

Specifies a rolling deployment strategy for updating a SageMaker AI inference component.

Contents

MaximumBatchSize

The batch size for each rolling step in the deployment process. For each step, SageMaker AI provisions capacity on the new endpoint fleet, routes traffic to that fleet, and terminates capacity on the old endpoint fleet. The value must be between 5% to 50% of the copy count of the inference component.

Type: InferenceComponentCapacitySize object

Required: Yes

WaitIntervalInSeconds

The length of the baking period, during which SageMaker AI monitors alarms for each batch on the new fleet.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 3600.

Required: Yes

MaximumExecutionTimeoutInSeconds

The time limit for the total deployment. Exceeding this limit causes a timeout.

Type: Integer

Valid Range: Minimum value of 600. Maximum value of 28800.

Required: No

RollbackMaximumBatchSize

The batch size for a rollback to the old endpoint fleet. If this field is absent, the value is set to the default, which is 100% of the total capacity. When the default is used, SageMaker AI provisions the entire capacity of the old fleet at once during rollback.

Type: InferenceComponentCapacitySize object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentRuntimeConfig

Service: Amazon SageMaker Service

Runtime settings for a model that is deployed with an inference component.

Contents

CopyCount

The number of runtime copies of the model container to deploy with the inference component. Each copy can serve inference requests.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentRuntimeConfigSummary

Service: Amazon SageMaker Service

Details about the runtime settings for the model that is deployed with the inference component.

Contents

CurrentCopyCount

The number of runtime copies of the model container that are currently deployed.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

DesiredCopyCount

The number of runtime copies of the model container that you requested to deploy with the inference component.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

PlacementStatus

The placement status of the inference component across instance types. Shows how the inference component copies are distributed across instance types.

Type: Array of InferenceComponentPlacementStatus objects

Array Members: Minimum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentSchedulingConfig

Service: Amazon SageMaker Service

The scheduling configuration that determines how inference component copies are placed across available instances when copies are added or removed.

Contents

PlacementStrategy

The strategy for placing inference component copies across available instances. If you also set `AvailabilityZoneBalance`, this strategy applies to placement within each Availability Zone.

SPREAD

Distributes copies evenly across available instances for better resilience.

BINPACK

Packs copies onto fewer instances to optimize resource utilization.

Type: String

Valid Values: `SPREAD | BINPACK`

Required: Yes

AvailabilityZoneBalance

Configuration for balancing inference component copies across Availability Zones.

Type: InferenceComponentAvailabilityZoneBalance object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentSpecification

Service: Amazon SageMaker Service

Details about the resources to deploy with this inference component, including the model, container, and compute resources.

Contents

BaseInferenceComponentName

The name of an existing inference component that is to contain the inference component that you're creating with your request.

Specify this parameter only if your request is meant to create an adapter inference component. An adapter inference component contains the path to an adapter model. The purpose of the adapter model is to tailor the inference output of a base foundation model, which is hosted by the base inference component. The adapter inference component uses the compute resources that you assigned to the base inference component.

When you create an adapter inference component, use the `Container` parameter to specify the location of the adapter artifacts. In the parameter value, use the `ArtifactUrl` parameter of the `InferenceComponentContainerSpecification` data type.

Before you can create an adapter inference component, you must have an existing inference component that contains the foundation model that you want to adapt.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

ComputeResourceRequirements

The compute resources allocated to run the model, plus any adapter models, that you assign to the inference component.

Omit this parameter if your request is meant to create an adapter inference component. An adapter inference component is loaded by a base inference component, and it uses the compute resources of the base inference component.

Type: InferenceComponentComputeResourceRequirements object

Required: No

Container

Defines a container that provides the runtime environment for a model that you deploy with an inference component.

Type: InferenceComponentContainerSpecification object

Required: No

DataCacheConfig

Settings that affect how the inference component caches data.

Type: InferenceComponentDataCacheConfig object

Required: No

InstanceType

The ML compute instance type for the inference component specification. Specifies which instance type this specification applies to. Required when using the `Specifications` parameter with multiple entries.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge`

| ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large | ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge | ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large | ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge | ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge | ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge | ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge | ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge | ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge | ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large | ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge | ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge | ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge | ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge | ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge

```
| ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge
| ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge |
ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge
| ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge |
ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge
| ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge |
ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge |
ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium
| ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge |
ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge
| ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge |
ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge
| ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge |
ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge |
ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large |
ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge
| ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge |
ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-
gb200.36xlarge | ml.p5.4xlarge
```

Required: No

ModelName

The name of an existing SageMaker AI model object in your account that you want to deploy with the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

SchedulingConfig

The scheduling configuration that determines how inference component copies are placed across available instances when copies are added or removed.

Type: InferenceComponentSchedulingConfig object

Required: No

StartupParameters

Settings that take effect while the model container starts up.

Type: InferenceComponentStartupParameters object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentSpecificationSummary

Service: Amazon SageMaker Service

Details about the resources that are deployed with this inference component.

Contents

BaseInferenceComponentName

The name of the base inference component that contains this inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

ComputeResourceRequirements

The compute resources allocated to run the model, plus any adapter models, that you assign to the inference component.

Type: InferenceComponentComputeResourceRequirements object

Required: No

Container

Details about the container that provides the runtime environment for the model that is deployed with the inference component.

Type: InferenceComponentContainerSpecificationSummary object

Required: No

DataCacheConfig

Settings that affect how the inference component caches data.

Type: InferenceComponentDataCacheConfigSummary object

Required: No

InstanceType

The ML compute instance type associated with this inference component specification.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large | ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge | ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge`

| ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large | ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge | ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge | ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge | ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge | ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge | ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge | ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large | ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge | ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge | ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge | ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge | ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium | ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge | ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge | ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge | ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge | ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge | ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge | ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large | ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge

`| ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge | ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge`

Required: No

ModelName

The name of the SageMaker AI model object that is deployed with the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

SchedulingConfig

The scheduling configuration that determines how inference component copies are placed across available instances when copies are added or removed.

Type: InferenceComponentSchedulingConfig object

Required: No

StartupParameters

Settings that take effect while the model container starts up.

Type: InferenceComponentStartupParameters object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentStartupParameters

Service: Amazon SageMaker Service

Settings that take effect while the model container starts up.

Contents

ContainerStartupHealthCheckTimeoutInSeconds

The timeout value, in seconds, for your inference container to pass health check by Amazon S3 Hosting. For more information about health check, see How Your Container Should Respond to Health Check (Ping) Requests.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 3600.

Required: No

ModelDataDownloadTimeoutInSeconds

The timeout value, in seconds, to download and extract the model that you want to host from Amazon S3 to the individual inference instance associated with this inference component.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 3600.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceComponentSummary

Service: Amazon SageMaker Service

A summary of the properties of an inference component.

Contents

CreationTime

The time when the inference component was created.

Type: Timestamp

Required: Yes

EndpointArn

The Amazon Resource Name (ARN) of the endpoint that hosts the inference component.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint/.*`

Required: Yes

EndpointName

The name of the endpoint that hosts the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

InferenceComponentArn

The Amazon Resource Name (ARN) of the inference component.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Required: Yes

InferenceComponentName

The name of the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

LastModifiedTime

The time when the inference component was last updated.

Type: Timestamp

Required: Yes

VariantName

The name of the production variant that hosts the inference component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

InferenceComponentStatus

The status of the inference component.

Type: String

Valid Values: `InService | Creating | Updating | Failed | Deleting`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceExecutionConfig

Service: Amazon SageMaker Service

Specifies details about how containers in a multi-container endpoint are run.

Contents

Mode

How containers in a multi-container are run. The following values are valid.

- `SERIAL` - Containers run as a serial pipeline.
- `DIRECT` - Only the individual container that you specify is run.

Type: String

Valid Values: `Serial | Direct`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceExperimentDataStorageConfig

Service: Amazon SageMaker Service

The Amazon S3 location and configuration for storing inference request and response data.

Contents

Destination

The Amazon S3 bucket where the inference request and response data is stored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `(https|s3)://([^/])/?(.*)`

Required: Yes

ContentType

Configuration specifying how to treat different headers. If no headers are specified Amazon SageMaker AI will by default base64 encode when capturing the data.

Type: CaptureContentTypeHeader object

Required: No

KmsKey

The AWS Key Management Service key that Amazon SageMaker uses to encrypt captured data at rest using Amazon S3 server-side encryption.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceExperimentSchedule

Service: Amazon SageMaker Service

The start and end times of an inference experiment.

The maximum duration that you can set for an inference experiment is 30 days.

Contents

EndTime

The timestamp at which the inference experiment ended or will end.

Type: Timestamp

Required: No

StartTime

The timestamp at which the inference experiment started or will start.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceExperimentSummary

Service: Amazon SageMaker Service

Lists a summary of properties of an inference experiment.

Contents

CreationTime

The timestamp at which the inference experiment was created.

Type: Timestamp

Required: Yes

LastModifiedTime

The timestamp when you last modified the inference experiment.

Type: Timestamp

Required: Yes

Name

The name of the inference experiment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: Yes

Status

The status of the inference experiment.

Type: String

Valid Values: `Creating | Created | Updating | Running | Starting | Stopping | Completed | Cancelled`

Required: Yes

Type

The type of the inference experiment.

Type: String

Valid Values: ShadowMode

Required: Yes

CompletionTime

The timestamp at which the inference experiment was completed.

Type: Timestamp

Required: No

Description

The description of the inference experiment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: .*

Required: No

RoleArn

The ARN of the IAM role that Amazon SageMaker can assume to access model artifacts and container images, and manage Amazon SageMaker Inference endpoints for model deployment.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+

Required: No

Schedule

The duration for which the inference experiment ran or will run.

The maximum duration that you can set for an inference experiment is 30 days.

Type: InferenceExperimentSchedule object

Required: No

StatusReason

The error message for the inference experiment status result.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceHubAccessConfig

Service: Amazon SageMaker Service

Configuration information specifying which hub contents have accessible deployment options.

Contents

HubContentArn

The ARN of the hub content for which deployment access is allowed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceMetrics

Service: Amazon SageMaker Service

The metrics for an existing endpoint compared in an Inference Recommender job.

Contents

MaxInvocations

The expected maximum number of requests per minute for the instance.

Type: Integer

Required: Yes

ModelLatency

The expected model latency at maximum invocations per minute for the instance.

Type: Integer

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceRecommendation

Service: Amazon SageMaker Service

A list of recommendations made by Amazon SageMaker Inference Recommender.

Contents

EndpointConfiguration

Defines the endpoint configuration parameters.

Type: EndpointOutputConfiguration object

Required: Yes

ModelConfiguration

Defines the model configuration.

Type: ModelConfiguration object

Required: Yes

InvocationEndTime

A timestamp that shows when the benchmark completed.

Type: Timestamp

Required: No

InvocationStartTime

A timestamp that shows when the benchmark started.

Type: Timestamp

Required: No

Metrics

The metrics used to decide what recommendation to make.

Type: RecommendationMetrics object

Required: No

RecommendationId

The recommendation ID which uniquely identifies each recommendation.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceRecommendationsJob

Service: Amazon SageMaker Service

A structure that contains a list of recommendation jobs.

Contents

CreationTime

A timestamp that shows when the job was created.

Type: Timestamp

Required: Yes

JobArn

The Amazon Resource Name (ARN) of the recommendation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:inference-recommendations-job/.*`

Required: Yes

JobDescription

The job description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Required: Yes

JobName

The name of the job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: Yes

JobType

The recommendation job type.

Type: String

Valid Values: `Default | Advanced`

Required: Yes

LastModifiedTime

A timestamp that shows when the job was last modified.

Type: Timestamp

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of an IAM role that enables Amazon SageMaker to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Status

The status of the job.

Type: String

Valid Values: `PENDING | IN_PROGRESS | COMPLETED | FAILED | STOPPING | STOPPED | DELETING | DELETED`

Required: Yes

CompletionTime

A timestamp that shows when the job completed.

Type: Timestamp

Required: No

FailureReason

If the job fails, provides information why the job failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

ModelName

The name of the created model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

ModelPackageVersionArn

The Amazon Resource Name (ARN) of a versioned model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: No

SamplePayloadUrl

The Amazon Simple Storage Service (Amazon S3) path where the sample payload is stored. This path must point to a single gzip compressed tar archive (.tar.gz suffix).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceRecommendationsJobStep

Service: Amazon SageMaker Service

A returned array object for the `Steps` response field in the ListInferenceRecommendationsJobSteps API command.

Contents

JobName

The name of the Inference Recommender job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: Yes

Status

The current status of the benchmark.

Type: String

Valid Values: `PENDING | IN_PROGRESS | COMPLETED | FAILED | STOPPING | STOPPED | DELETING | DELETED`

Required: Yes

StepType

The type of the subtask.

`BENCHMARK`: Evaluate the performance of your model on different instance types.

Type: String

Valid Values: `BENCHMARK`

Required: Yes

InferenceBenchmark

The details for a specific benchmark.

Type: RecommendationJobInferenceBenchmark object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceSpecification

Service: Amazon SageMaker Service

Defines how to perform inference generation after a training job is run.

Contents

Containers

The Amazon ECR registry path of the Docker image that contains the inference code.

Type: Array of ModelPackageContainerDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 15 items.

Required: Yes

SupportedContentTypes

The supported MIME types for the input data.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

SupportedRealtimeInferenceInstanceTypes

A list of the instance types that are used to generate inferences in real-time.

This parameter is required for unversioned models, and optional for versioned models.

Type: Array of strings

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge`

| ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large | ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge | ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large | ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge | ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge | ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge | ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge | ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge | ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge | ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large | ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge | ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge | ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge

| ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge | ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium | ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge | ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge | ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge | ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge | ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge | ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge | ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large | ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge | ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge | ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge

Required: No

SupportedResponseMIMETypes

The supported MIME types for the output data.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[-\w]+\/.+`

Required: No

SupportedTransformInstanceTypes

A list of the instance types on which a transformation job can be run or on which an endpoint can be deployed.

This parameter is required for unversioned models, and optional for versioned models.

Type: Array of strings

Array Members: Minimum number of 1 item.

Valid Values: `ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InfraCheckConfig

Service: Amazon SageMaker Service

Configuration information for the infrastructure health check of a training job. A SageMaker-provided health check tests the health of instance hardware and cluster network connectivity.

Contents

EnableInfraCheck

Enables an infrastructure health check.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InputConfig

Service: Amazon SageMaker Service

Contains information about the location of input model artifacts, the name and shape of the expected data inputs, and the framework in which the model was trained.

Contents

Framework

Identifies the framework in which the model was trained. For example: TENSORFLOW.

Type: String

Valid Values: `TENSORFLOW | KERAS | MXNET | ONNX | PYTORCH | XGBOOST | TFLITE | DARKNET | SKLEARN`

Required: Yes

S3Uri

The S3 path where the model artifacts, which result from model training, are stored. This path must point to a single gzip compressed tar archive (.tar.gz suffix).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

DataInputConfig

Specifies the name and shape of the expected data inputs for your trained model with a JSON dictionary form. The data inputs are `Framework` specific.

- `TensorFlow`: You must specify the name and shape (NHWC format) of the expected data inputs using a dictionary format for your trained model. The dictionary formats required for the console and CLI are different.
  - Examples for one input:
    - If using the console, `{"input":[1,1024,1024,3]}`

    - If using the CLI, `{\"input\":[1,1024,1024,3]}`
  - Examples for two inputs:
    - If using the console, `{"data1": [1,28,28,1], "data2":[1,28,28,1]}`
    - If using the CLI, `{\"data1\": [1,28,28,1], \"data2\":[1,28,28,1]}`
- `KERAS`: You must specify the name and shape (NCHW format) of expected data inputs using a dictionary format for your trained model. Note that while Keras model artifacts should be uploaded in NHWC (channel-last) format, `DataInputConfig` should be specified in NCHW (channel-first) format. The dictionary formats required for the console and CLI are different.
  - Examples for one input:
    - If using the console, `{"input_1":[1,3,224,224]}`
    - If using the CLI, `{\"input_1\":[1,3,224,224]}`
  - Examples for two inputs:
    - If using the console, `{"input_1": [1,3,224,224], "input_2":[1,3,224,224]}`
    - If using the CLI, `{\"input_1\": [1,3,224,224], \"input_2\":[1,3,224,224]}`
- `MXNET/ONNX/DARKNET`: You must specify the name and shape (NCHW format) of the expected data inputs in order using a dictionary format for your trained model. The dictionary formats required for the console and CLI are different.
  - Examples for one input:
    - If using the console, `{"data":[1,3,1024,1024]}`
    - If using the CLI, `{\"data\":[1,3,1024,1024]}`
  - Examples for two inputs:
    - If using the console, `{"var1": [1,1,28,28], "var2":[1,1,28,28]}`
    - If using the CLI, `{\"var1\": [1,1,28,28], \"var2\":[1,1,28,28]}`
- `PyTorch`: You can either specify the name and shape (NCHW format) of expected data inputs in order using a dictionary format for your trained model or you can specify the shape only using a list format. The dictionary formats required for the console and CLI are different. The list formats for the console and CLI are the same.
  - Examples for one input in dictionary format:
    - If using the console, `{"input0":[1,3,224,224]}`
    - If using the CLI, `{\"input0\":[1,3,224,224]}`
  - Example for one input in list format: `[[1,3,224,224]]`

  - Examples for two inputs in dictionary format:
    - If using the console, `{"input0":[1,3,224,224], "input1":[1,3,224,224]}`
    - If using the CLI, `{\"input0\":[1,3,224,224], \"input1\":[1,3,224,224]}`
  - Example for two inputs in list format: `[[1,3,224,224], [1,3,224,224]]`
- XGBOOST: input data name and shape are not needed.

`DataInputConfig` supports the following parameters for `CoreML` `TargetDevice` (ML Model format):

- `shape`: Input shape, for example `{"input_1": {"shape": [1,224,224,3]}}`. In addition to static input shapes, CoreML converter supports Flexible input shapes:
  - Range Dimension. You can use the Range Dimension feature if you know the input shape will be within some specific interval in that dimension, for example: `{"input_1": {"shape": ["1..10", 224, 224, 3]}}`
  - Enumerated shapes. Sometimes, the models are trained to work only on a select set of inputs. You can enumerate all supported input shapes, for example: `{"input_1": {"shape": [[1, 224, 224, 3], [1, 160, 160, 3]]}}`
- `default_shape`: Default input shape. You can set a default shape during conversion for both Range Dimension and Enumerated Shapes. For example `{"input_1": {"shape": ["1..10", 224, 224, 3], "default_shape": [1, 224, 224, 3]}}`
- `type`: Input type. Allowed values: `Image` and `Tensor`. By default, the converter generates an ML Model with inputs of type Tensor (MultiArray). User can set input type to be Image. Image input type requires additional input parameters such as `bias` and `scale`.
- `bias`: If the input type is an Image, you need to provide the bias vector.
- `scale`: If the input type is an Image, you need to provide a scale factor.

CoreML `ClassifierConfig` parameters can be specified using [OutputConfig](#) `CompilerOptions`. CoreML converter supports Tensorflow and PyTorch models. CoreML conversion examples:

- Tensor type input:
  - `"DataInputConfig": {"input_1": {"shape": [[1,224,224,3], [1,160,160,3]], "default_shape": [1,224,224,3]}}`
- Tensor type input without input name (PyTorch):
  - `"DataInputConfig": [{"shape": [[1,3,224,224], [1,3,160,160]], "default_shape": [1,3,224,224]}]`

- Image type input:
  - `"DataInputConfig": {"input_1": {"shape": [[1,224,224,3], [1,160,160,3]], "default_shape": [1,224,224,3], "type": "Image", "bias": [-1,-1,-1], "scale": 0.007843137255}}`
  - `"CompilerOptions": {"class_labels": "imagenet_labels_1000.txt"}`
- Image type input without input name (PyTorch):
  - `"DataInputConfig": [{"shape": [[1,3,224,224], [1,3,160,160]], "default_shape": [1,3,224,224], "type": "Image", "bias": [-1,-1,-1], "scale": 0.007843137255}]`
  - `"CompilerOptions": {"class_labels": "imagenet_labels_1000.txt"}`

Depending on the model format, `DataInputConfig` requires the following parameters for `ml_eia2` OutputConfig:TargetDevice.

- For TensorFlow models saved in the SavedModel format, specify the input names from `signature_def_key` and the input model shapes for `DataInputConfig`. Specify the `signature_def_key` in `OutputConfig:CompilerOptions` if the model does not use TensorFlow's default signature def key. For example:
  - `"DataInputConfig": {"inputs": [1, 224, 224, 3]}`
  - `"CompilerOptions": {"signature_def_key": "serving_custom"}`
- For TensorFlow models saved as a frozen graph, specify the input tensor names and shapes in `DataInputConfig` and the output tensor names for `output_names` in `OutputConfig:CompilerOptions`. For example:
  - `"DataInputConfig": {"input_tensor:0": [1, 224, 224, 3]}`
  - `"CompilerOptions": {"output_names": ["output_tensor:0"]}`

Type: String

Length Constraints: Minimum length of 1. Maximum length of 16384.

Pattern: `[\S\s]+`

Required: No

FrameworkVersion

Specifies the framework version to use. This API field is only supported for the MXNet, PyTorch, TensorFlow and TensorFlow Lite frameworks.

For information about framework versions supported for cloud targets and edge devices, see Cloud Supported Instance Types and Frameworks and Edge Supported Frameworks.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 10.

Pattern: `[0-9]\.[A-Za-z0-9.]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InstanceGroup

Service: Amazon SageMaker Service

Defines an instance group for heterogeneous cluster training. When requesting a training job using the CreateTrainingJob API, you can configure multiple instance groups .

Contents

InstanceCount

Specifies the number of instances of the instance group.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

InstanceGroupName

Specifies the name of the instance group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.+`

Required: Yes

InstanceType

Specifies the instance type of the instance group.

Type: String

Valid Values: `ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p3dn.24xlarge |`

ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5n.xlarge | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.8xlarge | ml.c6i.4xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.8xlarge | ml.r5d.12xlarge | ml.r5d.16xlarge | ml.r5d.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.p6-b200.48xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge | ml.p6-b300.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InstanceGroupHealthCheckConfiguration

Service: Amazon SageMaker Service

The configuration of deep health checks for an instance group.

> **Note**
>
> Overlapping deep health check configurations will be merged into a single operation.

Contents

DeepHealthChecks

A list of deep health checks to be performed.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `InstanceStress | InstanceConnectivity`

Required: Yes

InstanceGroupName

The name of the instance group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

InstanceIds

A list of Amazon Elastic Compute Cloud (EC2) instance IDs on which to perform deep health checks.

> **Note**
>
> Leave this field blank to perform deep health checks on the entire instance group.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 500 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `i-[a-f0-9]{8}(?:[a-f0-9]{9})?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InstanceGroupMetadata

Service: Amazon SageMaker Service

Metadata information about an instance group in a HyperPod cluster.

Contents

AmiOverride

If you use a custom Amazon Machine Image (AMI) for the instance group, this field shows the ID of the custom AMI.

Type: String

Required: No

AvailabilityZoneId

The ID of the Availability Zone where the instance group is located.

Type: String

Required: No

CapacityReservation

Information about the Capacity Reservation used by the instance group.

Type: CapacityReservation object

Required: No

FailureMessage

An error message describing why the instance group level operation (such as creating, scaling, or deleting) failed.

Type: String

Required: No

SecurityGroupIds

A list of security group IDs associated with the instance group.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: No

SubnetId

The ID of the subnet where the instance group is located.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InstanceGroupScalingMetadata

Service: Amazon SageMaker Service

Metadata information about scaling operations for an instance group.

Contents

FailureMessage

An error message describing why the scaling operation failed, if applicable.

Type: String

Required: No

InstanceCount

The current number of instances in the group.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

MinCount

Minimum instance count of the instance group.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

TargetCount

The desired number of instances for the group after scaling.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InstanceMetadata

Service: Amazon SageMaker Service

Metadata information about an instance in a HyperPod cluster.

Contents

AdditionalEnis

Information about additional Elastic Network Interfaces (ENIs) associated with the instance.

Type: AdditionalEnis object

Required: No

CapacityReservation

Information about the Capacity Reservation used by the instance.

Type: CapacityReservation object

Required: No

CustomerEni

The ID of the customer-managed Elastic Network Interface (ENI) associated with the instance.

Type: String

Required: No

FailureMessage

An error message describing why the instance creation or update failed, if applicable.

Type: String

Required: No

InstanceRequirementsEniConfigurations

The ENI configurations for the instance types in the instance requirements, grouped by network interface category (for example, ENI-only or EFA with ENIs). At most one configuration per category.

Type: Array of InstanceRequirementsEniConfiguration objects

Required: No

LcsExecutionState

The execution state of the Lifecycle Script (LCS) for the instance.

Type: String

Required: No

NodeLogicalId

The unique logical identifier of the node within the cluster. The ID used here is the same object as in the `BatchAddClusterNodes` API.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InstanceMetadataServiceConfiguration

Service: Amazon SageMaker Service

Information on the IMDS configuration of the notebook instance

Contents

MinimumInstanceMetadataServiceVersion

Indicates the minimum IMDS version that the notebook instance supports. When passed as part of `CreateNotebookInstance`, if no value is selected, then it defaults to IMDSv1. This means that both IMDSv1 and IMDSv2 are supported. If passed as part of `UpdateNotebookInstance`, there is no default.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1.

Pattern: 1 | 2

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InstancePlacementConfig

Service: Amazon SageMaker Service

Configuration for how instances are placed and allocated within UltraServers. This is only applicable for UltraServer capacity.

Contents

EnableMultipleJobs

If set to true, allows multiple jobs to share the same UltraServer instances. If set to false, ensures this job's instances are placed on an UltraServer exclusively, with no other jobs sharing the same UltraServer. Default is false.

Type: Boolean

Required: No

PlacementSpecifications

A list of specifications for how instances should be placed on specific UltraServers. Maximum of 10 items is supported.

Type: Array of PlacementSpecification objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InstancePool

Service: Amazon SageMaker Service

Specifies an instance type and its priority for a heterogeneous endpoint. Use instance pools to configure a production variant with multiple instance types, enabling the endpoint to provision instances across different types based on priority.

Contents

InstanceType

The ML compute instance type for the instance pool.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge`

```
| ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge
| ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge
| ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge
| ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge
| ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large |
ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge |
ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge
| ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge |
ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge
| ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge
| ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large
| ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge
| ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge |
ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge |
ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge |
ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge
| ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge |
ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large |
ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge |
ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge |
ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge
| ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge |
ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge |
ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge |
ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge
| ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge |
ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge
| ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge
| ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge |
ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge
| ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge |
ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge
| ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge |
ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge |
ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium
| ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge |
```

ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge | ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge | ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge | ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge | ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge | ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large | ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge | ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge | ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge

Required: Yes

Priority

The priority for the instance pool. SageMaker attempts to provision instances in order of priority, starting with the lowest value. If instances for a higher-priority pool are unavailable, SageMaker attempts to provision from the next pool.

Valid values: 1 to 5, where 1 is the highest priority.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 5.

Required: Yes

ModelNameOverride

The name of a SageMaker model to use for this instance pool instead of the model specified for the production variant. Use this to deploy a different model optimized for the instance type in this pool.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InstancePoolSummary

Service: Amazon SageMaker Service

A summary of an instance pool for a production variant, including the instance type and the current number of instances.

Contents

CurrentInstanceCount

The current number of instances of this type in the instance pool.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

InstanceType

The ML compute instance type for the instance pool.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large`

| ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large | ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge | ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large | ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge | ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge | ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge | ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge | ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge | ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge | ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large | ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge | ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge | ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge | ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge | ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge

`| ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium | ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge | ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge | ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge | ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge | ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge | ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge | ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large | ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge | ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge | ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InstanceRequirementsEniConfiguration

Service: Amazon SageMaker Service

The customer ENI and additional ENIs associated with a network interface category.

Contents

AdditionalEnis

Information about additional Elastic Network Interfaces (ENIs) associated with the instance type category.

Type: AdditionalEnis object

Required: No

CustomerEni

The ID of the customer-managed Elastic Network Interface (ENI) associated with the instance type category.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IntegerParameterRange

Service: Amazon SageMaker Service

For a hyperparameter of the integer type, specifies the range that a hyperparameter tuning job searches.

Contents

MaxValue

The maximum value of the hyperparameter to search.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

MinValue

The minimum value of the hyperparameter to search.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

Name

The name of the hyperparameter to search.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

ScalingType

The scale that hyperparameter tuning uses to search the hyperparameter range. For information about choosing a hyperparameter scale, see Hyperparameter Scaling. One of the following values:

Auto

SageMaker hyperparameter tuning chooses the best scale for the hyperparameter.

Linear

Hyperparameter tuning searches the values in the hyperparameter range by using a linear scale.

Logarithmic

Hyperparameter tuning searches the values in the hyperparameter range by using a logarithmic scale.

Logarithmic scaling works only for ranges that have only values greater than 0.

Type: String

Valid Values: `Auto | Linear | Logarithmic | ReverseLogarithmic`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IntegerParameterRangeSpecification

Service: Amazon SageMaker Service

Defines the possible values for an integer hyperparameter.

Contents

MaxValue

The maximum integer value allowed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

MinValue

The minimum integer value allowed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Job

Service: Amazon SageMaker Service

The properties of a job returned by the Search API.

Contents

CreationTime

The date and time that the job was created.

Type: Timestamp

Required: No

EndTime

The date and time that the job ended.

Type: Timestamp

Required: No

FailureReason

If the job failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

JobArn

The Amazon Resource Name (ARN) of the job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:(aws[a-z\-]*):sagemaker:[a-z0-9\-]+:[0-9]{12}:job/[a-zA-Z0-9]+/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

JobCategory

The category of the job.

Type: String

Valid Values: `AgentRFT | AgentRFTEvaluation`

Required: No

JobConfigDocument

The JSON configuration document for the job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 262144.

Required: No

JobConfigSchemaVersion

The schema version used for the job configuration document.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 16.

Pattern: `\d+\.\d+\.\d+`

Required: No

JobName

The name of the job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

JobStatus

The current status of the job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting | DeleteFailed`

Required: No

LastModifiedTime

The date and time that the job was last modified.

Type: Timestamp

Required: No

RoleArn

The ARN of the IAM role associated with the job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

SecondaryStatus

The detailed secondary status of the job, providing more granular information about the job's progress.

Type: String

Valid Values: `Starting | Downloading | Training | Uploading | Stopping | Stopped | MaxRuntimeExceeded | Interrupted | Failed | Completed | Restarting | Pending | Evaluating | Deleting | DeleteFailed`

Required: No

SecondaryStatusTransitions

A list of secondary status transitions for the job, with timestamps and optional status messages.

Type: Array of JobSecondaryStatusTransition objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Required: No

Tags

The tags associated with the job.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

JobConfigSchemaVersionSummary

Service: Amazon SageMaker Service

Provides summary information about a job configuration schema version.

Contents

JobConfigSchemaVersion

The version of the job configuration schema.

Type: String

Length Constraints: Minimum length of 5. Maximum length of 16.

Pattern: `\d+\.\d+\.\d+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

JobSecondaryStatusTransition

Service: Amazon SageMaker Service

Represents a secondary status transition for a job. Jobs progress through multiple secondary statuses during execution. Each transition records the status, start time, optional end time, and an optional message with additional details.

Contents

StartTime

The date and time that the status transition started.

Type: Timestamp

Required: Yes

Status

The secondary status of the job at this transition point.

Type: String

Valid Values: `Starting | Downloading | Training | Uploading | Stopping | Stopped | MaxRuntimeExceeded | Interrupted | Failed | Completed | Restarting | Pending | Evaluating | Deleting | DeleteFailed`

Required: Yes

EndTime

The date and time that the status transition ended.

Type: Timestamp

Required: No

StatusMessage

A detailed message about the status transition.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

JobStepMetadata

Service: Amazon SageMaker Service

Metadata for a SageMaker job step.

Contents

Arn

The Amazon Resource Name (ARN) of the SageMaker job that was run by this step execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

JobSummary

Service: Amazon SageMaker Service

Provides summary information about a job, returned by the `ListJobs` operation. Use `DescribeJob` to get full details for a specific job.

Contents

CreationTime

The date and time that the job was created.

Type: Timestamp

Required: Yes

JobArn

The Amazon Resource Name (ARN) of the job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:(aws[a-z\-]*):sagemaker:[a-z0-9\-]+:[0-9]{12}:job/[a-zA-Z0-9]+/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

JobCategory

The category of the job.

Type: String

Valid Values: `AgentRFT | AgentRFTEvaluation`

Required: Yes

JobName

The name of the job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

JobSecondaryStatus

The secondary status of the job, providing more granular information about the job's progress. Secondary statuses may change between releases.

Type: String

Valid Values: `Starting | Downloading | Training | Uploading | Stopping | Stopped | MaxRuntimeExceeded | Interrupted | Failed | Completed | Restarting | Pending | Evaluating | Deleting | DeleteFailed`

Required: Yes

JobStatus

The current status of the job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting | DeleteFailed`

Required: Yes

LastModifiedTime

The date and time that the job was last modified.

Type: Timestamp

Required: Yes

EndTime

The date and time that the job ended.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

JupyterLabAppImageConfig

Service: Amazon SageMaker Service

The configuration for the file system and kernels in a SageMaker AI image running as a JupyterLab app. The `FileSystemConfig` object is not supported.

Contents

ContainerConfig

The configuration used to run the application image container.

Type: ContainerConfig object

Required: No

FileSystemConfig

The Amazon Elastic File System storage configuration for a SageMaker AI image.

Type: FileSystemConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

JupyterLabAppSettings

Service: Amazon SageMaker Service

The settings for the JupyterLab application.

Contents

AppLifecycleManagement

Indicates whether idle shutdown is activated for JupyterLab applications.

Type: AppLifecycleManagement object

Required: No

BuiltInLifecycleConfigArn

The lifecycle configuration that runs before the default lifecycle configuration. It can override changes made in the default lifecycle configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:studio-lifecycle-config/.*|None)`

Required: No

CodeRepositories

A list of Git repositories that SageMaker automatically displays to users for cloning in the JupyterLab application.

Type: Array of CodeRepository objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

CustomImages

A list of custom SageMaker images that are configured to run as a JupyterLab app.

Type: Array of CustomImage objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

DefaultResourceSpec

Specifies the ARN's of a SageMaker AI image and SageMaker AI image version, and the instance type that the version runs on.

> **Note**
>
> When both `SageMakerImageVersionArn` and `SageMakerImageArn` are passed, `SageMakerImageVersionArn` is used. Any updates to `SageMakerImageArn` will not take effect if `SageMakerImageVersionArn` already exists in the `ResourceSpec` because `SageMakerImageVersionArn` always takes precedence. To clear the value set for `SageMakerImageVersionArn`, pass `None` as the value.

Type: ResourceSpec object

Required: No

EmrSettings

The configuration parameters that specify the IAM roles assumed by the execution role of SageMaker (assumable roles) and the cluster instances or job execution environments (execution roles or runtime roles) to manage and access resources required for running Amazon EMR clusters or Amazon EMR Serverless applications.

Type: EmrSettings object

Required: No

LifecycleConfigArns

The Amazon Resource Name (ARN) of the lifecycle configurations attached to the user profile or domain. To remove a lifecycle config, you must set `LifecycleConfigArns` to an empty list.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:studio-lifecycle-config/.*|None)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

JupyterServerAppSettings

Service: Amazon SageMaker Service

The JupyterServer app settings.

Contents

CodeRepositories

A list of Git repositories that SageMaker AI automatically displays to users for cloning in the JupyterServer application.

Type: Array of CodeRepository objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

DefaultResourceSpec

The default instance type and the Amazon Resource Name (ARN) of the default SageMaker AI image used by the JupyterServer app. If you use the `LifecycleConfigArns` parameter, then this parameter is also required.

Type: ResourceSpec object

Required: No

LifecycleConfigArns

The Amazon Resource Name (ARN) of the Lifecycle Configurations attached to the JupyterServerApp. If you use this parameter, the `DefaultResourceSpec` parameter is also required.

> **Note**
>
> To remove a Lifecycle Config, you must set `LifecycleConfigArns` to an empty list.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:studio-lifecycle-config/.*|None)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KendraSettings

Service: Amazon SageMaker Service

The Amazon SageMaker Canvas application setting where you configure document querying.

Contents

Status

Describes whether the document querying feature is enabled or disabled in the Canvas application.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KernelGatewayAppSettings

Service: Amazon SageMaker Service

The KernelGateway app settings.

Contents

CustomImages

A list of custom SageMaker AI images that are configured to run as a KernelGateway app.

The maximum number of custom images are as follows.

- On a domain level: 200
- On a space level: 5
- On a user profile level: 5

Type: Array of CustomImage objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

DefaultResourceSpec

The default instance type and the Amazon Resource Name (ARN) of the default SageMaker AI image used by the KernelGateway app.

> **Note**
>
> The Amazon SageMaker AI Studio UI does not use the default instance type value set here. The default instance type set here is used when Apps are created using the AWS CLI or CloudFormation and the instance type parameter value is not passed.

Type: ResourceSpec object

Required: No

LifecycleConfigArns

The Amazon Resource Name (ARN) of the Lifecycle Configurations attached to the the user profile or domain.

> **Note**
>
> To remove a Lifecycle Config, you must set `LifecycleConfigArns` to an empty list.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:studio-lifecycle-config/.*|None)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KernelGatewayImageConfig

Service: Amazon SageMaker Service

The configuration for the file system and kernels in a SageMaker AI image running as a KernelGateway app.

Contents

KernelSpecs

The specification of the Jupyter kernels in the image.

Type: Array of KernelSpec objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

FileSystemConfig

The Amazon Elastic File System storage configuration for a SageMaker AI image.

Type: FileSystemConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KernelSpec

Service: Amazon SageMaker Service

The specification of a Jupyter kernel.

Contents

Name

The name of the Jupyter kernel in the image. This value is case sensitive.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: Yes

DisplayName

The display name of the kernel.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelCounters

Service: Amazon SageMaker Service

Provides a breakdown of the number of objects labeled.

Contents

FailedNonRetryableError

The total number of objects that could not be labeled due to an error.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

HumanLabeled

The total number of objects labeled by a human worker.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

MachineLabeled

The total number of objects labeled by automated data labeling.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

TotalLabeled

The total number of objects labeled.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

Unlabeled

The total number of objects not yet labeled.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelCountersForWorkteam

Service: Amazon SageMaker Service

Provides counts for human-labeled tasks in the labeling job.

Contents

HumanLabeled

The total number of data objects labeled by a human worker.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

PendingHuman

The total number of data objects that need to be labeled by a human worker.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

Total

The total number of tasks in the labeling job.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobAlgorithmsConfig

Service: Amazon SageMaker Service

Provides configuration information for auto-labeling of your data objects. A `LabelingJobAlgorithmsConfig` object must be supplied in order to use auto-labeling.

Contents

LabelingJobAlgorithmSpecificationArn

Specifies the Amazon Resource Name (ARN) of the algorithm used for auto-labeling. You must select one of the following ARNs:

- *Image classification*

  `arn:aws:sagemaker:`*`region`*`:027400017018:labeling-job-algorithm-specification/image-classification`

- *Text classification*

  `arn:aws:sagemaker:`*`region`*`:027400017018:labeling-job-algorithm-specification/text-classification`

- *Object detection*

  `arn:aws:sagemaker:`*`region`*`:027400017018:labeling-job-algorithm-specification/object-detection`

- *Semantic Segmentation*

  `arn:aws:sagemaker:`*`region`*`:027400017018:labeling-job-algorithm-specification/semantic-segmentation`

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:.*`

Required: Yes

InitialActiveLearningModelArn

At the end of an auto-label job Ground Truth sends the Amazon Resource Name (ARN) of the final model used for auto-labeling. You can use this model as the starting point for subsequent similar jobs by providing the ARN of the model here.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:model/.*`

Required: No

LabelingJobResourceConfig

Provides configuration information for a labeling job.

Type: LabelingJobResourceConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobDataAttributes

Service: Amazon SageMaker Service

Attributes of the data specified by the customer. Use these to describe the data to be labeled.

Contents

ContentClassifiers

Declares that your content is free of personally identifiable information or adult content. SageMaker may restrict the Amazon Mechanical Turk workers that can view your task based on this information.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 256 items.

Valid Values: `FreeOfPersonallyIdentifiableInformation | FreeOfAdultContent`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobDataSource

Service: Amazon SageMaker Service

Provides information about the location of input data.

You must specify at least one of the following: `S3DataSource` or `SnsDataSource`.

Use `SnsDataSource` to specify an SNS input topic for a streaming labeling job. If you do not specify and SNS input topic ARN, Ground Truth will create a one-time labeling job.

Use `S3DataSource` to specify an input manifest file for both streaming and one-time labeling jobs. Adding an `S3DataSource` is optional if you use `SnsDataSource` to create a streaming labeling job.

Contents

S3DataSource

The Amazon S3 location of the input data objects.

Type: LabelingJobS3DataSource object

Required: No

SnsDataSource

An Amazon SNS data source used for streaming labeling jobs. To learn more, see Send Data to a Streaming Labeling Job.

Type: LabelingJobSnsDataSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobForWorkteamSummary

Service: Amazon SageMaker Service

Provides summary information for a work team.

Contents

CreationTime

The date and time that the labeling job was created.

Type: Timestamp

Required: Yes

JobReferenceCode

A unique identifier for a labeling job. You can use this to refer to a specific labeling job.

Type: String

Length Constraints: Minimum length of 1.

Pattern: `.+`

Required: Yes

WorkRequesterAccountId

The AWS account ID of the account used to start the labeling job.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `\d+`

Required: Yes

LabelCounters

Provides information about the progress of a labeling job.

Type: LabelCountersForWorkteam object

Required: No

LabelingJobName

The name of the labeling job that the work team is assigned to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

NumberOfHumanWorkersPerDataObject

The configured number of workers per data object.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 9.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobInputConfig

Service: Amazon SageMaker Service

Input configuration information for a labeling job.

Contents

DataSource

The location of the input data.

Type: LabelingJobDataSource object

Required: Yes

DataAttributes

Attributes of the data specified by the customer.

Type: LabelingJobDataAttributes object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobOutput

Service: Amazon SageMaker Service

Specifies the location of the output produced by the labeling job.

Contents

OutputDatasetS3Uri

The Amazon S3 bucket location of the manifest file for labeled data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

FinalActiveLearningModelArn

The Amazon Resource Name (ARN) for the most recent SageMaker model trained as part of automated data labeling.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:model/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobOutputConfig

Service: Amazon SageMaker Service

Output configuration information for a labeling job.

Contents

S3OutputPath

The Amazon S3 location to write output data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

KmsKeyId

The AWS Key Management Service ID of the key used to encrypt the output data, if any.

If you provide your own KMS key ID, you must add the required permissions to your KMS key described in Encrypt Output Data and Storage Volume with AWS KMS.

If you don't provide a KMS key ID, Amazon SageMaker uses the default AWS KMS key for Amazon S3 for your role's account to encrypt your output data.

If you use a bucket policy with an `s3:PutObject` permission that only allows objects with server-side encryption, set the condition key of `s3:x-amz-server-side-encryption` to `"aws:kms"`. For more information, see KMS-Managed Encryption Keys in the *Amazon Simple Storage Service Developer Guide.*

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

SnsTopicArn

An Amazon Simple Notification Service (Amazon SNS) output topic ARN. Provide a `SnsTopicArn` if you want to do real time chaining to another streaming job and receive an Amazon SNS notifications each time a data object is submitted by a worker.

If you provide an `SnsTopicArn` in `OutputConfig`, when workers complete labeling tasks, Ground Truth will send labeling task output data to the SNS output topic you specify here.

To learn more, see Receive Output Data from a Streaming Labeling Job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sns:[a-z0-9\-]*:[0-9]{12}:[a-zA-Z0-9_.-]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobResourceConfig

Service: Amazon SageMaker Service

Configure encryption on the storage volume attached to the ML compute instance used to run automated data labeling model training and inference.

Contents

VolumeKmsKeyId

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt data on the storage volume attached to the ML compute instance(s) that run the training and inference jobs used for automated data labeling.

You can only specify a `VolumeKmsKeyId` when you create a labeling job with automated data labeling enabled using the API operation `CreateLabelingJob`. You cannot specify an AWS KMS key to encrypt the storage volume used for automated data labeling model training and inference when you create a labeling job using the console. To learn more, see Output Data and Storage Volume Encryption.

The `VolumeKmsKeyId` can be any of the following formats:

- KMS Key ID

  ```
  "1234abcd-12ab-34cd-56ef-1234567890ab"
  ```

- Amazon Resource Name (ARN) of a KMS Key

  ```
  "arn:aws:kms:us-
  west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab"
  ```

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

VpcConfig

Specifies an Amazon Virtual Private Cloud (VPC) that your SageMaker jobs, hosted models, and compute resources have access to. You can control access to and from your resources by

configuring a VPC. For more information, see Give SageMaker Access to Resources in your Amazon VPC.

Type: VpcConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobS3DataSource

Service: Amazon SageMaker Service

The Amazon S3 location of the input data objects.

Contents

ManifestS3Uri

The Amazon S3 location of the manifest file that describes the input data objects.

The input manifest file referenced in `ManifestS3Uri` must contain one of the following keys: `source-ref` or `source`. The value of the keys are interpreted as follows:

- `source-ref`: The source of the object is the Amazon S3 object specified in the value. Use this value when the object is a binary object, such as an image.
- `source`: The source of the object is the value. Use this value when the object is a text value.

If you are a new user of Ground Truth, it is recommended you review Use an Input Manifest File in the Amazon SageMaker Developer Guide to learn how to create an input manifest file.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobSnsDataSource

Service: Amazon SageMaker Service

An Amazon SNS data source used for streaming labeling jobs.

Contents

SnsTopicArn

The Amazon SNS input topic Amazon Resource Name (ARN). Specify the ARN of the input topic you will use to send new data objects to a streaming labeling job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sns:[a-z0-9\-]*:[0-9]{12}:[a-zA-Z0-9_.-]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobStoppingConditions

Service: Amazon SageMaker Service

A set of conditions for stopping a labeling job. If any of the conditions are met, the job is automatically stopped. You can use these conditions to control the cost of data labeling.

> **Note**
>
> Labeling jobs fail after 30 days with an appropriate client error message.

Contents

MaxHumanLabeledObjectCount

The maximum number of objects that can be labeled by human workers.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

MaxPercentageOfInputDatasetLabeled

The maximum number of input data objects that should be labeled.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LabelingJobSummary

Service: Amazon SageMaker Service

Provides summary information about a labeling job.

Contents

CreationTime

The date and time that the job was created (timestamp).

Type: Timestamp

Required: Yes

LabelCounters

Counts showing the progress of the labeling job.

Type: LabelCounters object

Required: Yes

LabelingJobArn

The Amazon Resource Name (ARN) assigned to the labeling job when it was created.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:labeling-job/.*`

Required: Yes

LabelingJobName

The name of the labeling job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

LabelingJobStatus

The current status of the labeling job.

Type: String

Valid Values: `Initializing | InProgress | Completed | Failed | Stopping | Stopped`

Required: Yes

LastModifiedTime

The date and time that the job was last modified (timestamp).

Type: Timestamp

Required: Yes

WorkteamArn

The Amazon Resource Name (ARN) of the work team assigned to the job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:workteam/.*`

Required: Yes

AnnotationConsolidationLambdaArn

The Amazon Resource Name (ARN) of the Lambda function used to consolidate the annotations from individual workers into a label for a data object. For more information, see Annotation Consolidation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:lambda:[a-z0-9\-]*:[0-9]{12}:function:.*`

Required: No

FailureReason

If the `LabelingJobStatus` field is `Failed`, this field contains a description of the error.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

InputConfig

Input configuration for the labeling job.

Type: LabelingJobInputConfig object

Required: No

LabelingJobOutput

The location of the output produced by the labeling job.

Type: LabelingJobOutput object

Required: No

PreHumanTaskLambdaArn

The Amazon Resource Name (ARN) of a Lambda function. The function is run before each data object is sent to a worker.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:lambda:[a-z0-9\-]*:[0-9]{12}:function:.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

LambdaStepMetadata

Service: Amazon SageMaker Service

Metadata for a Lambda step.

Contents

Arn

The Amazon Resource Name (ARN) of the Lambda function that was run by this step execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

OutputParameters

A list of the output parameters of the Lambda step.

Type: Array of OutputParameter objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LastUpdateStatus

Service: Amazon SageMaker Service

A value that indicates whether the update was successful.

Contents

Status

A value that indicates whether the update was made successful.

Type: String

Valid Values: `Successful | Failed | InProgress`

Required: Yes

FailureReason

If the update wasn't successful, indicates the reason why it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LineageGroupSummary

Service: Amazon SageMaker Service

Lists a summary of the properties of a lineage group. A lineage group provides a group of shareable lineage entity resources.

Contents

CreationTime

The creation time of the lineage group summary.

Type: Timestamp

Required: No

DisplayName

The display name of the lineage group summary.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

LastModifiedTime

The last modified time of the lineage group summary.

Type: Timestamp

Required: No

LineageGroupArn

The Amazon Resource Name (ARN) of the lineage group resource.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:lineage-group/.*`

Required: No

LineageGroupName

The name or Amazon Resource Name (ARN) of the lineage group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LineageMetadata

Service: Amazon SageMaker Service

The metadata that tracks relationships between ML artifacts, actions, and contexts.

Contents

ActionArns

The Amazon Resource Name (ARN) of the lineage action.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 5 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2048.

Value Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

ArtifactArns

The Amazon Resource Name (ARN) of the lineage artifact.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 5 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2048.

Value Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

Associations

The lineage associations.

Type: Array of AssociationInfo objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

ContextArns

The Amazon Resource Name (ARN) of the lineage context.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 5 items.

Key Length Constraints: Minimum length of 0. Maximum length of 2048.

Value Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ManagedConfiguration

Service: Amazon SageMaker Service

The managed configuration of a model package group.

Contents

ManagedStorageType

The storage type of the model package.

Type: String

Valid Values: `Restricted`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MemberDefinition

Service: Amazon SageMaker Service

Defines an Amazon Cognito or your own OIDC IdP user group that is part of a work team.

Contents

CognitoMemberDefinition

The Amazon Cognito user group that is part of the work team.

Type: CognitoMemberDefinition object

Required: No

OidcMemberDefinition

A list user groups that exist in your OIDC Identity Provider (IdP). One to ten groups can be used to create a single private work team. When you add a user group to the list of `Groups`, you can add that user group to one or more private work teams. If you add a user group to a private work team, all workers in that user group are added to the work team.

Type: OidcMemberDefinition object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetadataProperties

Service: Amazon SageMaker Service

Metadata properties of the tracking entity, trial, or trial component.

Contents

CommitId

The commit ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

GeneratedBy

The entity this entity was generated by.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

ProjectId

The project ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

Repository

The repository.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetricData

Service: Amazon SageMaker Service

The name, value, and date and time of a metric that was emitted to Amazon CloudWatch.

Contents

MetricName

The name of the metric.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: No

Timestamp

The date and time that the algorithm emitted the metric.

Type: Timestamp

Required: No

Value

The value of the metric.

Type: Float

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetricDatum

Service: Amazon SageMaker Service

Information about the metric for a candidate produced by an AutoML job.

Contents

MetricName

The name of the metric.

Type: String

Valid Values: `Accuracy | MSE | F1 | F1macro | AUC | RMSE | BalancedAccuracy | R2 | Recall | RecallMacro | Precision | PrecisionMacro | MAE | MAPE | MASE | WAPE | AverageWeightedQuantileLoss`

Required: No

Set

The dataset split from which the AutoML job produced the metric.

Type: String

Valid Values: `Train | Validation | Test`

Required: No

StandardMetricName

The name of the standard metric.

> **Note**
>
> For definitions of the standard metrics, see `Autopilot candidate metrics`.

Type: String

Valid Values: `Accuracy | MSE | F1 | F1macro | AUC | RMSE | MAE | R2 | BalancedAccuracy | Precision | PrecisionMacro | Recall | RecallMacro | LogLoss | InferenceLatency | MAPE | MASE | WAPE |`

`AverageWeightedQuantileLoss | Rouge1 | Rouge2 | RougeL | RougeLSum | Perplexity | ValidationLoss | TrainingLoss`

Required: No

Value

The value of the metric.

Type: Float

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetricDefinition

Service: Amazon SageMaker Service

Specifies a metric that the training algorithm writes to `stderr` or `stdout`. You can view these logs to understand how your training job performs and check for any errors encountered during training. SageMaker hyperparameter tuning captures all defined metrics. Specify one of the defined metrics to use as an objective metric using the TuningObjective parameter in the `HyperParameterTrainingJobDefinition` API to evaluate job performance during hyperparameter tuning.

Contents

Name

The name of the metric.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: Yes

Regex

A regular expression that searches the output of a training job and gets the value of the metric. For more information about using regular expressions to define metrics, see Defining metrics and environment variables.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Pattern: `.+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetricsConfig

Service: Amazon SageMaker Service

The configuration for Utilization metrics.

Contents

EnableDetailedObservability

Indicates whether detailed observability is enabled for the endpoint. When set to `True`, the following metrics are published at the configured frequency:

- Container-level inference metrics scraped from the container's Prometheus endpoint (such as request latency, error counts, and throughput). Available metrics vary by framework.
- Per-GPU metrics (utilization, memory, and temperature) attributed to individual inference components.
- Per-instance host metrics (CPU, memory, and disk utilization).
- Inference component placement metrics (copy count per Availability Zone).

For first-party and Deep Learning Containers (DLC), the Prometheus endpoint path is determined automatically. For Bring-Your-Own-Container (BYOC) cases, you can optionally set `ContainerMetricsConfig` to specify a custom endpoint path. If not specified, the default path `/metrics` on port `8080` is used.

When set to `False`, these additional metrics are not published. Standard invocation and utilization metrics controlled by `EnableEnhancedMetrics` are unaffected.

The default value for new endpoint configurations is `True`. For existing endpoint configurations created before this feature, the value is `False` unless explicitly set.

Type: Boolean

Required: No

EnableEnhancedMetrics

Specifies whether to enable enhanced metrics for the endpoint. Enhanced metrics provide utilization and invocation data at instance and container granularity. Container granularity is supported for Inference Components. The default is `False`.

Type: Boolean

Required: No

MetricPublishFrequencyInSeconds

The interval, in seconds, at which metrics are published to Amazon CloudWatch. Defaults to `60`. Valid values: `10`, `30`, `60`, `120`, `180`, `240`, `300`.

When `EnableEnhancedMetrics` is set to `False`, this interval applies to utilization metrics only. Invocation metrics continue to be published at the default 60-second interval. When `EnableEnhancedMetrics` is set to `True`, this interval applies to both utilization and invocation metrics.

When `EnableDetailedObservability` is set to `True`, this interval applies to per-GPU metrics, per-instance host metrics, container metrics, and fleet-level inference component lifecycle and placement metrics.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetricsEndpoint

Service: Amazon SageMaker Service

Specifies a metrics endpoint for a container, including the path where the container exposes Prometheus-formatted metrics and the frequency at which to publish them to Amazon CloudWatch.

Contents

MetricsEndpointPath

The path to the metrics endpoint exposed by the container. For example, `/metrics` or `/server/metrics`. The path must start with `/` and can contain alphanumeric characters, forward slashes, underscores, hyphens, and periods. Maximum length is 256 characters. If not specified, defaults to `/metrics`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `/(?!.*\.\.)[a-zA-Z0-9/_.\-]+`

Required: Yes

MetricPublishFrequencyInSeconds

The interval, in seconds, at which container metrics scraped from the endpoint are published to Amazon CloudWatch. Valid values: `10`, `30`, `60`, `120`, `180`, `240`, `300`. Defaults to `60`.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetricSpecification

Service: Amazon SageMaker Service

An object containing information about a metric.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

Customized

Information about a customized metric.

Type: CustomizedMetricSpecification object

Required: No

Predefined

Information about a predefined metric.

Type: PredefinedMetricSpecification object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetricsSource

Service: Amazon SageMaker Service

Details about the metrics source.

Contents

ContentType

The metric source content type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

S3Uri

The S3 URI for the metrics source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

ContentDigest

The hash key used for the metrics source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 72.

Pattern: `[Ss][Hh][Aa]256:[0-9a-fA-F]{64}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MlflowAppSummary

Service: Amazon SageMaker Service

The summary of the Mlflow App to list.

Contents

Arn

The ARN of a listed MLflow App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-app/.*`

Required: No

CreationTime

The creation time of a listed MLflow App.

Type: Timestamp

Required: No

LastModifiedTime

The last modified time of a listed MLflow App.

Type: Timestamp

Required: No

MlflowVersion

The version of a listed MLflow App.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 16.

Pattern: `[0-9]*.[0-9]*.[0-9]*`

Required: No

Name

The name of the MLflow App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: No

Status

The status of the MLflow App.

Type: String

Valid Values: `Creating | Created | CreateFailed | Updating | Updated | UpdateFailed | Deleting | DeleteFailed | Deleted`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MlflowConfig

Service: Amazon SageMaker Service

The MLflow configuration using SageMaker managed MLflow.

Contents

MlflowResourceArn

The Amazon Resource Name (ARN) of the MLflow resource.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-[a-zA-Z-]*/.*`

Required: Yes

MlflowExperimentName

The MLflow experiment name used for this job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*`

Required: No

MlflowRunName

The MLflow run name used for this job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MLflowConfiguration

Service: Amazon SageMaker Service

The MLflow configuration.

Contents

MlflowExperimentName

The name of the MLflow configuration.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*`

Required: No

MlflowResourceArn

The Amazon Resource Name (ARN) of MLflow configuration resource.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-[a-zA-Z-]*/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MlflowDetails

Service: Amazon SageMaker Service

The MLflow details of this job.

Contents

MlflowExperimentId

The MLflow experiment ID used for this job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*`

Required: No

MlflowRunId

The MLflow run ID used for this job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Model

Service: Amazon SageMaker Service

The properties of a model as returned by the Search API.

Contents

Containers

The containers in the inference pipeline.

Type: Array of ContainerDefinition objects

Array Members: Minimum number of 0 items. Maximum number of 15 items.

Required: No

CreationTime

A timestamp that indicates when the model was created.

Type: Timestamp

Required: No

DeploymentRecommendation

A set of recommended deployment configurations for the model.

Type: DeploymentRecommendation object

Required: No

EnableNetworkIsolation

Isolates the model container. No inbound or outbound network calls can be made to or from the model container.

Type: Boolean

Required: No

ExecutionRoleArn

The Amazon Resource Name (ARN) of the IAM role that you specified for the model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

InferenceExecutionConfig

Specifies details about how containers in a multi-container endpoint are run.

Type: InferenceExecutionConfig object

Required: No

ModelArn

The Amazon Resource Name (ARN) of the model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:model/.*`

Required: No

ModelName

The name of the model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

PrimaryContainer

Describes the container, as part of model definition.

Type: ContainerDefinition object

Required: No

Tags

A list of key-value pairs associated with the model. For more information, see Tagging AWS resources in the *AWS General Reference Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

VpcConfig

Specifies an Amazon Virtual Private Cloud (VPC) that your SageMaker jobs, hosted models, and compute resources have access to. You can control access to and from your resources by configuring a VPC. For more information, see Give SageMaker Access to Resources in your Amazon VPC.

Type: VpcConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelAccessConfig

Service: Amazon SageMaker Service

The access configuration file to control access to the ML model. You can explicitly accept the model end-user license agreement (EULA) within the `ModelAccessConfig`.

- If you are a Jumpstart user, see the End-user license agreements section for more details on accepting the EULA.
- If you are an AutoML user, see the *Optional Parameters* section of *Create an AutoML job to fine-tune text generation models using the API* for details on How to set the EULA acceptance when fine-tuning a model using the AutoML API.

Contents

AcceptEula

Specifies agreement to the model end-user license agreement (EULA). The `AcceptEula` value must be explicitly defined as `True` in order to accept the EULA that this model requires. You are responsible for reviewing and complying with any applicable license terms and making sure they are acceptable for your use case before downloading or using a model.

Type: Boolean

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelArtifacts

Service: Amazon SageMaker Service

Provides information about the location that is configured for storing model artifacts.

Model artifacts are outputs that result from training a model. They typically consist of trained parameters, a model definition that describes how to compute inferences, and other metadata. A SageMaker container stores your trained model artifacts in the `/opt/ml/model` directory. After training has completed, by default, these artifacts are uploaded to your Amazon S3 bucket as compressed files.

Contents

S3ModelArtifacts

The path of the S3 object that contains the model artifacts. For example, `s3://bucket-name/keynameprefix/model.tar.gz`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelBiasAppSpecification

Service: Amazon SageMaker Service

Docker container image configuration object for the model bias job.

Contents

ConfigUri

JSON formatted S3 file that defines bias parameters. For more information on this JSON configuration file, see Configure bias parameters.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

ImageUri

The container image to be run by the model bias job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

Environment

Sets the environment variables in the Docker container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[\S\s]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelBiasBaselineConfig

Service: Amazon SageMaker Service

The configuration for a baseline model bias job.

Contents

BaseliningJobName

The name of the baseline model bias job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ConstraintsResource

The constraints resource for a monitoring job.

Type: MonitoringConstraintsResource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelBiasJobInput

Service: Amazon SageMaker Service

Inputs for the model bias job.

Contents

GroundTruthS3Input

Location of ground truth labels to use in model bias job.

Type: MonitoringGroundTruthS3Input object

Required: Yes

BatchTransformInput

Input object for the batch transform job.

Type: BatchTransformInput object

Required: No

EndpointInput

Input object for the endpoint

Type: EndpointInput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelCard

Service: Amazon SageMaker Service

An Amazon SageMaker Model Card.

Contents

Content

The content of the model card. Content uses the model card JSON schema and provided as a string.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 100000.

Pattern: `.*`

Required: No

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

CreationTime

The date and time that the model card was created.

Type: Timestamp

Required: No

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

LastModifiedTime

The date and time that the model card was last modified.

Type: Timestamp

Required: No

ModelCardArn

The Amazon Resource Name (ARN) of the model card.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-card/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelCardName

The unique name of the model card.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelCardStatus

The approval status of the model card within your organization. Different organizations might have different criteria for model card review and approval.

- `Draft`: The model card is a work in progress.
- `PendingReview`: The model card is pending review.
- `Approved`: The model card is approved.
- `Archived`: The model card is archived. No more updates should be made to the model card, but it can still be exported.

Type: String

Valid Values: `Draft | PendingReview | Approved | Archived`

Required: No

ModelCardVersion

The version of the model card.

Type: Integer

Required: No

ModelId

The unique name (ID) of the model.

Type: String

Required: No

ModelPackageGroupName

The model package group that contains the model package. Only relevant for model cards created for model packages in the Amazon SageMaker Model Registry.

Type: String

Required: No

RiskRating

The risk rating of the model. Different organizations might have different criteria for model card risk ratings. For more information, see Risk ratings.

Type: String

Required: No

SecurityConfig

The security configuration used to protect model card data.

Type: ModelCardSecurityConfig object

Required: No

Tags

Key-value pairs used to manage metadata for the model card.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelCardExportArtifacts

Service: Amazon SageMaker Service

The artifacts of the model card export job.

Contents

S3ExportArtifacts

The Amazon S3 URI of the exported model artifacts.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelCardExportJobSummary

Service: Amazon SageMaker Service

The summary of the Amazon SageMaker Model Card export job.

Contents

CreatedAt

The date and time that the model card export job was created.

Type: Timestamp

Required: Yes

LastModifiedAt

The date and time that the model card export job was last modified..

Type: Timestamp

Required: Yes

ModelCardExportJobArn

The Amazon Resource Name (ARN) of the model card export job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-card/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}/export-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelCardExportJobName

The name of the model card export job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelCardName

The name of the model card that the export job exports.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelCardVersion

The version of the model card that the export job exports.

Type: Integer

Required: Yes

Status

The completion status of the model card export job.

Type: String

Valid Values: `InProgress | Completed | Failed`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelCardExportOutputConfig

Service: Amazon SageMaker Service

Configure the export output details for an Amazon SageMaker Model Card.

Contents

S3OutputPath

The Amazon S3 output path to export your model card PDF.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelCardSecurityConfig

Service: Amazon SageMaker Service

Configure the security settings to protect model card data.

Contents

KmsKeyId

A Key Management Service key ID to use for encrypting a model card.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelCardSummary

Service: Amazon SageMaker Service

A summary of the model card.

Contents

CreationTime

The date and time that the model card was created.

Type: Timestamp

Required: Yes

ModelCardArn

The Amazon Resource Name (ARN) of the model card.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-card/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelCardName

The name of the model card.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelCardStatus

The approval status of the model card within your organization. Different organizations might have different criteria for model card review and approval.

- `Draft`: The model card is a work in progress.
- `PendingReview`: The model card is pending review.
- `Approved`: The model card is approved.
- `Archived`: The model card is archived. No more updates should be made to the model card, but it can still be exported.

Type: String

Valid Values: `Draft | PendingReview | Approved | Archived`

Required: Yes

LastModifiedTime

The date and time that the model card was last modified.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelCardVersionSummary

Service: Amazon SageMaker Service

A summary of a specific version of the model card.

Contents

CreationTime

The date and time that the model card version was created.

Type: Timestamp

Required: Yes

ModelCardArn

The Amazon Resource Name (ARN) of the model card.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-card/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelCardName

The name of the model card.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelCardStatus

The approval status of the model card version within your organization. Different organizations might have different criteria for model card review and approval.

- `Draft`: The model card is a work in progress.
- `PendingReview`: The model card is pending review.
- `Approved`: The model card is approved.
- `Archived`: The model card is archived. No more updates should be made to the model card, but it can still be exported.

Type: String

Valid Values: `Draft | PendingReview | Approved | Archived`

Required: Yes

ModelCardVersion

A version of the model card.

Type: Integer

Required: Yes

LastModifiedTime

The time date and time that the model card version was last modified.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelClientConfig

Service: Amazon SageMaker Service

Configures the timeout and maximum number of retries for processing a transform job invocation.

Contents

InvocationsMaxRetries

The maximum number of retries when invocation requests are failing. The default value is 3.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 3.

Required: No

InvocationsTimeoutInSeconds

The timeout value in seconds for an invocation request. The default value is 600.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 3600.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelCompilationConfig

Service: Amazon SageMaker Service

Settings for the model compilation technique that's applied by a model optimization job.

Contents

Image

The URI of an LMI DLC in Amazon ECR. SageMaker uses this image to run the optimization.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[\S]+`

Required: No

OverrideEnvironment

Environment variables that override the default ones in the model container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 25 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `(?!\s*$).+`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ModelConfiguration

Service: Amazon SageMaker Service

Defines the model configuration. Includes the specification name and environment parameters.

Contents

CompilationJobName

The name of the compilation job used to create the recommended model artifacts.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

EnvironmentParameters

Defines the environment parameters that includes key, value types, and values.

Type: Array of EnvironmentParameter objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

InferenceSpecificationName

The inference specification name in the model package version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelDashboardEndpoint

Service: Amazon SageMaker Service

An endpoint that hosts a model displayed in the Amazon SageMaker Model Dashboard.

Contents

CreationTime

A timestamp that indicates when the endpoint was created.

Type: Timestamp

Required: Yes

EndpointArn

The Amazon Resource Name (ARN) of the endpoint.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:endpoint/.*`

Required: Yes

EndpointName

The endpoint name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

EndpointStatus

The endpoint status.

Type: String

Valid Values: `OutOfService | Creating | Updating | SystemUpdating | RollingBack | InService | Deleting | Failed | UpdateRollbackFailed`

Required: Yes

LastModifiedTime

The last time the endpoint was modified.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelDashboardIndicatorAction

Service: Amazon SageMaker Service

An alert action taken to light up an icon on the Amazon SageMaker Model Dashboard when an alert goes into `InAlert` status.

Contents

Enabled

Indicates whether the alert action is turned on.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelDashboardModel

Service: Amazon SageMaker Service

A model displayed in the Amazon SageMaker Model Dashboard.

Contents

Endpoints

The endpoints that host a model.

Type: Array of ModelDashboardEndpoint objects

Required: No

LastBatchTransformJob

A batch transform job. For information about SageMaker batch transform, see Use Batch Transform.

Type: TransformJob object

Required: No

Model

A model displayed in the Model Dashboard.

Type: Model object

Required: No

ModelCard

The model card for a model.

Type: ModelDashboardModelCard object

Required: No

MonitoringSchedules

The monitoring schedules for a model.

Type: Array of ModelDashboardMonitoringSchedule objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelDashboardModelCard

Service: Amazon SageMaker Service

The model card for a model displayed in the Amazon SageMaker Model Dashboard.

Contents

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

CreationTime

A timestamp that indicates when the model card was created.

Type: Timestamp

Required: No

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

LastModifiedTime

A timestamp that indicates when the model card was last updated.

Type: Timestamp

Required: No

ModelCardArn

The Amazon Resource Name (ARN) for a model card.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-card/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelCardName

The name of a model card.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelCardStatus

The model card status.

Type: String

Valid Values: `Draft | PendingReview | Approved | Archived`

Required: No

ModelCardVersion

The model card version.

Type: Integer

Required: No

ModelId

For models created in SageMaker, this is the model ARN. For models created outside of SageMaker, this is a user-customized string.

Type: String

Required: No

RiskRating

A model card's risk rating. Can be low, medium, or high.

Type: String

Required: No

SecurityConfig

The KMS Key ID (`KMSKeyId`) for encryption of model card information.

Type: ModelCardSecurityConfig object

Required: No

Tags

The tags associated with a model card.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelDashboardMonitoringSchedule

Service: Amazon SageMaker Service

A monitoring schedule for a model displayed in the Amazon SageMaker Model Dashboard.

Contents

BatchTransformInput

Input object for the batch transform job.

Type: BatchTransformInput object

Required: No

CreationTime

A timestamp that indicates when the monitoring schedule was created.

Type: Timestamp

Required: No

EndpointName

The endpoint which is monitored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

FailureReason

If a monitoring job failed, provides the reason.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

LastModifiedTime

A timestamp that indicates when the monitoring schedule was last updated.

Type: Timestamp

Required: No

LastMonitoringExecutionSummary

Summary of information about the last monitoring job to run.

Type: MonitoringExecutionSummary object

Required: No

MonitoringAlertSummaries

A JSON array where each element is a summary for a monitoring alert.

Type: Array of MonitoringAlertSummary objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

MonitoringScheduleArn

The Amazon Resource Name (ARN) of a monitoring schedule.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

MonitoringScheduleConfig

Configures the monitoring schedule and defines the monitoring job.

Type: MonitoringScheduleConfig object

Required: No

MonitoringScheduleName

The name of a monitoring schedule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MonitoringScheduleStatus

The status of the monitoring schedule.

Type: String

Valid Values: `Pending | Failed | Scheduled | Stopped`

Required: No

MonitoringType

The monitor type of a model monitor.

Type: String

Valid Values: `DataQuality | ModelQuality | ModelBias | ModelExplainability`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelDataQuality

Service: Amazon SageMaker Service

Data quality constraints and statistics for a model.

Contents

Constraints

Data quality constraints for a model.

Type: MetricsSource object

Required: No

Statistics

Data quality statistics for a model.

Type: MetricsSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelDataSource

Service: Amazon SageMaker Service

Specifies the location of ML model data to deploy. If specified, you must specify one and only one of the available data sources.

Contents

S3DataSource

Specifies the S3 location of ML model data to deploy.

Type: S3ModelDataSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelDeployConfig

Service: Amazon SageMaker Service

Specifies how to generate the endpoint name for an automatic one-click Autopilot model deployment.

Contents

AutoGenerateEndpointName

Set to `True` to automatically generate an endpoint name for a one-click Autopilot model deployment; set to `False` otherwise. The default value is `False`.

> **Note**
>
> If you set `AutoGenerateEndpointName` to `True`, do not specify the `EndpointName`; otherwise a 400 error is thrown.

Type: Boolean

Required: No

EndpointName

Specifies the endpoint name to use for a one-click Autopilot model deployment if the endpoint name is not generated automatically.

> **Note**
>
> Specify the `EndpointName` if and only if you set `AutoGenerateEndpointName` to `False`; otherwise a 400 error is thrown.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelDeployResult

Service: Amazon SageMaker Service

Provides information about the endpoint of the model deployment.

Contents

EndpointName

The name of the endpoint to which the model has been deployed.

> **Note**
>
> If model deployment fails, this field is omitted from the response.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelDigests

Service: Amazon SageMaker Service

Provides information to verify the integrity of stored model artifacts.

Contents

ArtifactDigest

Provides a hash value that uniquely identifies the stored model artifacts.

Type: String

Pattern: `[\w:]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelExplainabilityAppSpecification

Service: Amazon SageMaker Service

Docker container image configuration object for the model explainability job.

Contents

ConfigUri

JSON formatted Amazon S3 file that defines explainability parameters. For more information on this JSON configuration file, see Configure model explainability parameters.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

ImageUri

The container image to be run by the model explainability job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

Environment

Sets the environment variables in the Docker container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[\S\s]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelExplainabilityBaselineConfig

Service: Amazon SageMaker Service

The configuration for a baseline model explainability job.

Contents

BaseliningJobName

The name of the baseline model explainability job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ConstraintsResource

The constraints resource for a monitoring job.

Type: MonitoringConstraintsResource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelExplainabilityJobInput

Service: Amazon SageMaker Service

Inputs for the model explainability job.

Contents

BatchTransformInput

Input object for the batch transform job.

Type: BatchTransformInput object

Required: No

EndpointInput

Input object for the endpoint

Type: EndpointInput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelInfrastructureConfig

Service: Amazon SageMaker Service

The configuration for the infrastructure that the model will be deployed to.

Contents

InfrastructureType

The inference option to which to deploy your model. Possible values are the following:

- `RealTime`: Deploy to real-time inference.

Type: String

Valid Values: `RealTimeInference`

Required: Yes

RealTimeInferenceConfig

The infrastructure configuration for deploying the model to real-time inference.

Type: RealTimeInferenceConfig object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelInput

Service: Amazon SageMaker Service

Input object for the model.

Contents

DataInputConfig

The input configuration object for the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 16384.

Pattern: `[\S\s]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelLatencyThreshold

Service: Amazon SageMaker Service

The model latency threshold.

Contents

Percentile

The model latency percentile threshold. Acceptable values are P95 and P99. For custom load tests, specify the value as P95.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Required: No

ValueInMilliseconds

The model latency percentile value in milliseconds.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelLifeCycle

Service: Amazon SageMaker Service

A structure describing the current state of the model in its life cycle.

Contents

Stage

The current stage in the model life cycle.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

StageStatus

The current status of a stage in model life cycle.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

StageDescription

Describes the stage related details.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.{0,1024}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelMetadataFilter

Service: Amazon SageMaker Service

Part of the search expression. You can specify the name and value (domain, task, framework, framework version, task, and model).

Contents

Name

The name of the of the model to filter by.

Type: String

Valid Values: `Domain | Framework | Task | FrameworkVersion`

Required: Yes

Value

The value to filter the model metadata.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelMetadataSearchExpression

Service: Amazon SageMaker Service

One or more filters that searches for the specified resource or resources in a search. All resource objects that satisfy the expression's condition are included in the search results

Contents

Filters

A list of filter objects.

Type: Array of ModelMetadataFilter objects

Array Members: Minimum number of 1 item. Maximum number of 4 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelMetadataSummary

Service: Amazon SageMaker Service

A summary of the model metadata.

Contents

Domain

The machine learning domain of the model.

Type: String

Required: Yes

Framework

The machine learning framework of the model.

Type: String

Required: Yes

FrameworkVersion

The framework version of the model.

Type: String

Required: Yes

Model

The name of the model.

Type: String

Required: Yes

Task

The machine learning task of the model.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelMetrics

Service: Amazon SageMaker Service

Contains metrics captured from a model.

Contents

Bias

Metrics that measure bias in a model.

Type: Bias object

Required: No

Explainability

Metrics that help explain a model.

Type: Explainability object

Required: No

ModelDataQuality

Metrics that measure the quality of the input data for a model.

Type: ModelDataQuality object

Required: No

ModelQuality

Metrics that measure the quality of a model.

Type: ModelQuality object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackage

Service: Amazon SageMaker Service

A container for your trained model that can be deployed for SageMaker inference. This can include inference code, artifacts, and metadata. The model package type can be one of the following.

- Versioned model: A part of a model package group in Model Registry.
- Unversioned model: Not part of a model package group and used in AWS Marketplace.

For more information, see `CreateModelPackage`.

Contents

AdditionalInferenceSpecifications

An array of additional Inference Specification objects.

Type: Array of AdditionalInferenceSpecificationDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 15 items.

Required: No

ApprovalDescription

A description provided when the model approval is set.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

CertifyForMarketplace

Whether the model package is to be certified to be listed on AWS Marketplace. For information about listing model packages on AWS Marketplace, see List Your Algorithm or Model Package on AWS Marketplace.

Type: Boolean

Required: No

CreatedBy

Information about the user who created or modified an experiment, trial, trial component, lineage group, or project.

Type: UserContext object

Required: No

CreationTime

The time that the model package was created.

Type: Timestamp

Required: No

CustomerMetadataProperties

The metadata properties for the model package.

Type: String to string map

Map Entries: Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `([\p{L}\p{Z}\p{N}_.:\/=+\-@]*)${1,128}`

Value Length Constraints: Minimum length of 1. Maximum length of 256.

Value Pattern: `([\p{L}\p{Z}\p{N}_.:\/=+\-@]*)${1,256}`

Required: No

Domain

The machine learning domain of your model package and its components. Common machine learning domains include computer vision and natural language processing.

Type: String

Required: No

DriftCheckBaselines

Represents the drift check baselines that can be used when the model monitor is set using the model package.

Type: DriftCheckBaselines object

Required: No

InferenceSpecification

Defines how to perform inference generation after a training job is run.

Type: InferenceSpecification object

Required: No

LastModifiedBy

Information about the user who created or modified an experiment, trial, trial component, lineage group, or project.

Type: UserContext object

Required: No

LastModifiedTime

The last time the model package was modified.

Type: Timestamp

Required: No

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Required: No

ModelApprovalStatus

The approval status of the model. This can be one of the following values.

- APPROVED - The model is approved

- `REJECTED` - The model is rejected.
- `PENDING_MANUAL_APPROVAL` - The model is waiting for manual approval.

Type: String

Valid Values: `Approved | Rejected | PendingManualApproval`

Required: No

ModelCard

The model card associated with the model package. Since `ModelPackageModelCard` is tied to a model package, it is a specific usage of a model card and its schema is simplified compared to the schema of `ModelCard`. The `ModelPackageModelCard` schema does not include `model_package_details`, and `model_overview` is composed of the `model_creator` and `model_artifact` properties. For more information about the model package model card schema, see Model package model card schema. For more information about the model card associated with the model package, see View the Details of a Model Version.

Type: ModelPackageModelCard object

Required: No

ModelLifeCycle

A structure describing the current state of the model in its life cycle.

Type: ModelLifeCycle object

Required: No

ModelMetrics

Metrics for the model.

Type: ModelMetrics object

Required: No

ModelPackageArn

The Amazon Resource Name (ARN) of the model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: No

ModelPackageDescription

The description of the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

ModelPackageGroupName

The model group to which the model belongs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelPackageName

The name of the model package. The name can be as follows:

- For a versioned model, the name is automatically generated by SageMaker Model Registry and follows the format '`ModelPackageGroupName/ModelPackageVersion`'.
- For an unversioned model, you must provide the name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelPackageRegistrationType

The package registration type of the model package.

Type: String

Valid Values: `Logged | Registered`

Required: No

ModelPackageStatus

The status of the model package. This can be one of the following values.

- `PENDING` - The model package is pending being created.
- `IN_PROGRESS` - The model package is in the process of being created.
- `COMPLETED` - The model package was successfully created.
- `FAILED` - The model package failed.
- `DELETING` - The model package is in the process of being deleted.

Type: String

Valid Values: `Pending | InProgress | Completed | Failed | Deleting`

Required: No

ModelPackageStatusDetails

Specifies the validation and image scan statuses of the model package.

Type: ModelPackageStatusDetails object

Required: No

ModelPackageVersion

The version number of a versioned model.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

SamplePayloadUrl

The Amazon Simple Storage Service path where the sample payload are stored. This path must point to a single gzip compressed tar archive (.tar.gz suffix).

Type: String

Required: No

SecurityConfig

An optional AWS Key Management Service key to encrypt, decrypt, and re-encrypt model package information for regulated workloads with highly sensitive data.

Type: ModelPackageSecurityConfig object

Required: No

SkipModelValidation

Indicates if you want to skip model validation.

Type: String

Valid Values: `All | None`

Required: No

SourceAlgorithmSpecification

A list of algorithms that were used to create a model package.

Type: SourceAlgorithmSpecification object

Required: No

SourceUri

The URI of the source for the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{N}\p{P}]{0,1024}`

Required: No

Tags

A list of the tags associated with the model package. For more information, see Tagging AWS resources in the *AWS General Reference Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

Task

The machine learning task your model package accomplishes. Common machine learning tasks include object detection and image classification.

Type: String

Required: No

ValidationSpecification

Specifies batch transform jobs that SageMaker runs to validate your model package.

Type: ModelPackageValidationSpecification object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackageConfig

Service: Amazon SageMaker Service

The configuration for the Model package.

Contents

ModelPackageGroupArn

The Amazon Resource Name (ARN) of the model package group of output model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package-group/[\S]{1,2048}`

Required: Yes

SourceModelPackageArn

The Amazon Resource Name (ARN) of the source model package used for continued fine-tuning and custom model evaluation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ModelPackageContainerDefinition

Service: Amazon SageMaker Service

Describes the Docker container for the model package.

Contents

AdditionalModelDataSources

Data sources that are available to your model in addition to the one that you specify for `ModelDataSource` when you use the `CreateModelPackage` action.

Type: Array of AdditionalModelDataSource objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

AdditionalS3DataSource

The additional data source that is used during inference in the Docker container for your model package.

Type: AdditionalS3DataSource object

Required: No

BaseModel

Identifies the foundation model that was used as the starting point for model customization.

Type: BaseModel object

Required: No

ContainerHostname

The DNS host name for the Docker container.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

Environment

The environment variables to set in the Docker container. Each key and value in the `Environment` string to string map can have length of up to 1024. We support up to 16 entries in the map.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 1024.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 1024.

Value Pattern: `[\S\s]*`

Required: No

Framework

The machine learning framework of the model package container image.

Type: String

Required: No

FrameworkVersion

The framework version of the Model Package Container Image.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 10.

Pattern: `[0-9]\.[A-Za-z0-9.-]+`

Required: No

Image

The Amazon Elastic Container Registry (Amazon ECR) path where inference code is stored.

If you are using your own custom algorithm instead of an algorithm provided by SageMaker, the inference code must meet SageMaker requirements. SageMaker supports both `registry/repository[:tag]` and `registry/repository[@digest]` image path formats. For more information, see Using Your Own Algorithms with Amazon SageMaker.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[\S]+`

Required: No

ImageDigest

An MD5 hash of the training algorithm that identifies the Docker image used for training.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 72.

Pattern: `[Ss][Hh][Aa]256:[0-9a-fA-F]{64}`

Required: No

IsCheckpoint

Specifies whether the model data is a training checkpoint.

Type: Boolean

Required: No

ModelDataETag

The ETag associated with Model Data URL.

Type: String

Required: No

ModelDataSource

Specifies the location of ML model data to deploy during endpoint creation.

Type: ModelDataSource object

Required: No

ModelDataUrl

The Amazon S3 path where the model artifacts, which result from model training, are stored. This path must point to a single `gzip` compressed tar archive (`.tar.gz` suffix).

> **Note**
>
> The model artifacts must be in an S3 bucket that is in the same region as the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

ModelInput

A structure with Model Input details.

Type: ModelInput object

Required: No

NearestModelName

The name of a pre-trained machine learning benchmarked by Amazon SageMaker Inference Recommender model that matches your model. You can find a list of benchmarked models by calling `ListModelMetadata`.

Type: String

Required: No

ProductId

The AWS Marketplace product ID of the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackageGroup

Service: Amazon SageMaker Service

A group of versioned models in the Model Registry.

Contents

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

CreationTime

The time that the model group was created.

Type: Timestamp

Required: No

ModelPackageGroupArn

The Amazon Resource Name (ARN) of the model group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package-group/[\S]{1,2048}`

Required: No

ModelPackageGroupDescription

The description for the model group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

ModelPackageGroupName

The name of the model group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelPackageGroupStatus

The status of the model group. This can be one of the following values.

- `PENDING` - The model group is pending being created.
- `IN_PROGRESS` - The model group is in the process of being created.
- `COMPLETED` - The model group was successfully created.
- `FAILED` - The model group failed.
- `DELETING` - The model group is in the process of being deleted.
- `DELETE_FAILED` - SageMaker failed to delete the model group.

Type: String

Valid Values: `Pending | InProgress | Completed | Failed | Deleting | DeleteFailed`

Required: No

Tags

A list of the tags associated with the model group. For more information, see Tagging AWS resources in the *AWS General Reference Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackageGroupSummary

Service: Amazon SageMaker Service

Summary information about a model group.

Contents

CreationTime

The time that the model group was created.

Type: Timestamp

Required: Yes

ModelPackageGroupArn

The Amazon Resource Name (ARN) of the model group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package-group/[\S]{1,2048}`

Required: Yes

ModelPackageGroupName

The name of the model group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ModelPackageGroupStatus

The status of the model group.

Type: String

Valid Values: `Pending | InProgress | Completed | Failed | Deleting | DeleteFailed`

Required: Yes

ManagedConfiguration

The managed configuration of the model package group.

Type: ManagedConfiguration object

Required: No

ModelPackageGroupDescription

A description of the model group.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackageModelCard

Service: Amazon SageMaker Service

The model card associated with the model package. Since `ModelPackageModelCard` is tied to a model package, it is a specific usage of a model card and its schema is simplified compared to the schema of `ModelCard`. The `ModelPackageModelCard` schema does not include `model_package_details`, and `model_overview` is composed of the `model_creator` and `model_artifact` properties. For more information about the model package model card schema, see Model package model card schema. For more information about the model card associated with the model package, see View the Details of a Model Version.

Contents

ModelCardContent

The content of the model card. The content must follow the schema described in Model Package Model Card Schema.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 100000.

Pattern: `.*`

Required: No

ModelCardStatus

The approval status of the model card within your organization. Different organizations might have different criteria for model card review and approval.

- `Draft`: The model card is a work in progress.
- `PendingReview`: The model card is pending review.
- `Approved`: The model card is approved.
- `Archived`: The model card is archived. No more updates can be made to the model card content. If you try to update the model card content, you will receive the message `Model Card is in Archived state`.

Type: String

Valid Values: `Draft | PendingReview | Approved | Archived`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackageSecurityConfig

Service: Amazon SageMaker Service

An optional AWS Key Management Service key to encrypt, decrypt, and re-encrypt model package information for regulated workloads with highly sensitive data.

Contents

KmsKeyId

The AWS KMS Key ID (`KMSKeyId`) used for encryption of model package information.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackageStatusDetails

Service: Amazon SageMaker Service

Specifies the validation and image scan statuses of the model package.

Contents

ValidationStatuses

The validation status of the model package.

Type: Array of ModelPackageStatusItem objects

Required: Yes

ImageScanStatuses

The status of the scan of the Docker image container for the model package.

Type: Array of ModelPackageStatusItem objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackageStatusItem

Service: Amazon SageMaker Service

Represents the overall status of a model package.

Contents

Name

The name of the model package for which the overall status is being reported.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

Status

The current status.

Type: String

Valid Values: `NotStarted | InProgress | Completed | Failed`

Required: Yes

FailureReason

if the overall status is `Failed`, the reason for the failure.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackageSummary

Service: Amazon SageMaker Service

Provides summary information about a model package.

Contents

CreationTime

A timestamp that shows when the model package was created.

Type: Timestamp

Required: Yes

ModelPackageArn

The Amazon Resource Name (ARN) of the model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: Yes

ModelPackageStatus

The overall status of the model package.

Type: String

Valid Values: `Pending | InProgress | Completed | Failed | Deleting`

Required: Yes

ModelApprovalStatus

The approval status of the model. This can be one of the following values.

- `APPROVED` - The model is approved
- `REJECTED` - The model is rejected.
- `PENDING_MANUAL_APPROVAL` - The model is waiting for manual approval.

Type: String

Valid Values: `Approved | Rejected | PendingManualApproval`

Required: No

ModelLifeCycle

A structure describing the current state of the model in its life cycle.

Type: ModelLifeCycle object

Required: No

ModelPackageDescription

A brief description of the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

ModelPackageGroupName

If the model package is a versioned model, the model group that the versioned model belongs to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelPackageName

The name of the model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ModelPackageRegistrationType

The package registration type of the model package summary.

Type: String

Valid Values: `Logged | Registered`

Required: No

ModelPackageVersion

If the model package is a versioned model, the version of the model.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackageValidationProfile

Service: Amazon SageMaker Service

Contains data, such as the inputs and targeted instance types that are used in the process of validating the model package.

The data provided in the validation profile is made available to your buyers on AWS Marketplace.

Contents

ProfileName

The name of the profile for the model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

TransformJobDefinition

The `TransformJobDefinition` object that describes the transform job used for the validation of the model package.

Type: TransformJobDefinition object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackageValidationSpecification

Service: Amazon SageMaker Service

Specifies batch transform jobs that SageMaker runs to validate your model package.

Contents

ValidationProfiles

An array of `ModelPackageValidationProfile` objects, each of which specifies a batch transform job that SageMaker runs to validate your model package.

Type: Array of ModelPackageValidationProfile objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: Yes

ValidationRole

The IAM roles to be used for the validation of the model package.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelQuality

Service: Amazon SageMaker Service

Model quality statistics and constraints.

Contents

Constraints

Model quality constraints.

Type: MetricsSource object

Required: No

Statistics

Model quality statistics.

Type: MetricsSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelQualityAppSpecification

Service: Amazon SageMaker Service

Container image configuration object for the monitoring job.

Contents

ImageUri

The address of the container image that the monitoring job runs.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

ContainerArguments

An array of arguments for the container used to run the monitoring job.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

ContainerEntrypoint

Specifies the entrypoint for a container that the monitoring job runs.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

Environment

Sets the environment variables in the container that the monitoring job runs.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[\S\s]*`

Required: No

PostAnalyticsProcessorSourceUri

An Amazon S3 URI to a script that is called after analysis has been performed. Applicable only for the built-in (first party) containers.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

ProblemType

The machine learning problem type of the model that the monitoring job monitors.

Type: String

Valid Values: `BinaryClassification | MulticlassClassification | Regression`

Required: No

RecordPreprocessorSourceUri

An Amazon S3 URI to a script that is called per row prior to running analysis. It can base64 decode the payload and convert it into a flattened JSON so that the built-in container can use the converted data. Applicable only for the built-in (first party) containers.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelQualityBaselineConfig

Service: Amazon SageMaker Service

Configuration for monitoring constraints and monitoring statistics. These baseline resources are compared against the results of the current job from the series of jobs scheduled to collect data periodically.

Contents

BaseliningJobName

The name of the job that performs baselining for the monitoring job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ConstraintsResource

The constraints resource for a monitoring job.

Type: MonitoringConstraintsResource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelQualityJobInput

Service: Amazon SageMaker Service

The input for the model quality monitoring job. Currently endpoints are supported for input for model quality monitoring jobs.

Contents

GroundTruthS3Input

The ground truth label provided for the model.

Type: MonitoringGroundTruthS3Input object

Required: Yes

BatchTransformInput

Input object for the batch transform job.

Type: BatchTransformInput object

Required: No

EndpointInput

Input object for the endpoint

Type: EndpointInput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelQuantizationConfig

Service: Amazon SageMaker Service

Settings for the model quantization technique that's applied by a model optimization job.

Contents

Image

The URI of an LMI DLC in Amazon ECR. SageMaker uses this image to run the optimization.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[\S]+`

Required: No

OverrideEnvironment

Environment variables that override the default ones in the model container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 25 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `(?!\s*$).+`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ModelRegisterSettings

Service: Amazon SageMaker Service

The model registry settings for the SageMaker Canvas application.

Contents

CrossAccountModelRegisterRoleArn

The Amazon Resource Name (ARN) of the SageMaker model registry account. Required only to register model versions created by a different SageMaker Canvas AWS account than the AWS account in which SageMaker model registry is set up.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

Status

Describes whether the integration to the model registry is enabled or disabled in the Canvas application.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelShardingConfig

Service: Amazon SageMaker Service

Settings for the model sharding technique that's applied by a model optimization job.

Contents

Image

The URI of an LMI DLC in Amazon ECR. SageMaker uses this image to run the optimization.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[\S]+`

Required: No

OverrideEnvironment

Environment variables that override the default ones in the model container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 25 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `(?!\s*$).+`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ModelSpeculativeDecodingConfig

Service: Amazon SageMaker Service

Settings for the model speculative decoding technique that's applied by a model optimization job.

Contents

Technique

The speculative decoding technique to apply during model optimization.

Type: String

Valid Values: EAGLE

Required: Yes

TrainingDataSource

The location of the training data to use for speculative decoding. The data must be formatted as ShareGPT, OpenAI Completions or OpenAI Chat Completions. The input can also be unencrypted captured data from a SageMaker endpoint as long as the endpoint uses one of the above formats.

Type: ModelSpeculativeDecodingTrainingDataSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelSpeculativeDecodingTrainingDataSource

Service: Amazon SageMaker Service

Contains information about the training data source for speculative decoding.

Contents

S3DataType

The type of data stored in the Amazon S3 location. Valid values are `S3Prefix` or `ManifestFile`.

Type: String

Valid Values: `S3Prefix | ManifestFile`

Required: Yes

S3Uri

The Amazon S3 URI that points to the training data for speculative decoding.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelStepMetadata

Service: Amazon SageMaker Service

Metadata for Model steps.

Contents

Arn

The Amazon Resource Name (ARN) of the created model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelSummary

Service: Amazon SageMaker Service

Provides summary information about a model.

Contents

CreationTime

A timestamp that indicates when the model was created.

Type: Timestamp

Required: Yes

ModelArn

The Amazon Resource Name (ARN) of the model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:model/.*`

Required: Yes

ModelName

The name of the model that you want a summary for.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelVariantConfig

Service: Amazon SageMaker Service

Contains information about the deployment options of a model.

Contents

InfrastructureConfig

The configuration for the infrastructure that the model will be deployed to.

Type: ModelInfrastructureConfig object

Required: Yes

ModelName

The name of the Amazon SageMaker Model entity.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

VariantName

The name of the variant.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelVariantConfigSummary

Service: Amazon SageMaker Service

Summary of the deployment configuration of a model.

Contents

InfrastructureConfig

The configuration of the infrastructure that the model has been deployed to.

Type: ModelInfrastructureConfig object

Required: Yes

ModelName

The name of the Amazon SageMaker Model entity.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

Status

The status of deployment for the model variant on the hosted inference endpoint.

- `Creating` - Amazon SageMaker is preparing the model variant on the hosted inference endpoint.
- `InService` - The model variant is running on the hosted inference endpoint.
- `Updating` - Amazon SageMaker is updating the model variant on the hosted inference endpoint.
- `Deleting` - Amazon SageMaker is deleting the model variant on the hosted inference endpoint.
- `Deleted` - The model variant has been deleted on the hosted inference endpoint. This can only happen after stopping the experiment.

Type: String

Valid Values: `Creating | Updating | InService | Deleting | Deleted`

Required: Yes

VariantName

The name of the variant.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringAlertActions

Service: Amazon SageMaker Service

A list of alert actions taken in response to an alert going into `InAlert` status.

Contents

ModelDashboardIndicator

An alert action taken to light up an icon on the Model Dashboard when an alert goes into `InAlert` status.

Type: ModelDashboardIndicatorAction object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringAlertHistorySummary

Service: Amazon SageMaker Service

Provides summary information of an alert's history.

Contents

AlertStatus

The current alert status of an alert.

Type: String

Valid Values: `InAlert | OK`

Required: Yes

CreationTime

A timestamp that indicates when the first alert transition occurred in an alert history. An alert transition can be from status `InAlert` to `OK`, or from `OK` to `InAlert`.

Type: Timestamp

Required: Yes

MonitoringAlertName

The name of a monitoring alert.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

MonitoringScheduleName

The name of a monitoring schedule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringAlertSummary

Service: Amazon SageMaker Service

Provides summary information about a monitor alert.

Contents

Actions

A list of alert actions taken in response to an alert going into `InAlert` status.

Type: MonitoringAlertActions object

Required: Yes

AlertStatus

The current status of an alert.

Type: String

Valid Values: `InAlert | OK`

Required: Yes

CreationTime

A timestamp that indicates when a monitor alert was created.

Type: Timestamp

Required: Yes

DatapointsToAlert

Within `EvaluationPeriod`, how many execution failures will raise an alert.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: Yes

EvaluationPeriod

The number of most recent monitoring executions to consider when evaluating alert status.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: Yes

LastModifiedTime

A timestamp that indicates when a monitor alert was last updated.

Type: Timestamp

Required: Yes

MonitoringAlertName

The name of a monitoring alert.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringAppSpecification

Service: Amazon SageMaker Service

Container image configuration object for the monitoring job.

Contents

ImageUri

The container image to be run by the monitoring job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

ContainerArguments

An array of arguments for the container used to run the monitoring job.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 50 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

ContainerEntrypoint

Specifies the entrypoint for a container used to run the monitoring job.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

PostAnalyticsProcessorSourceUri

An Amazon S3 URI to a script that is called after analysis has been performed. Applicable only for the built-in (first party) containers.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

RecordPreprocessorSourceUri

An Amazon S3 URI to a script that is called per row prior to running analysis. It can base64 decode the payload and convert it into a flattened JSON so that the built-in container can use the converted data. Applicable only for the built-in (first party) containers.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringBaselineConfig

Service: Amazon SageMaker Service

Configuration for monitoring constraints and monitoring statistics. These baseline resources are compared against the results of the current job from the series of jobs scheduled to collect data periodically.

Contents

BaseliningJobName

The name of the job that performs baselining for the monitoring job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ConstraintsResource

The baseline constraint file in Amazon S3 that the current monitoring job should validated against.

Type: MonitoringConstraintsResource object

Required: No

StatisticsResource

The baseline statistics file in Amazon S3 that the current monitoring job should be validated against.

Type: MonitoringStatisticsResource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringClusterConfig

Service: Amazon SageMaker Service

Configuration for the cluster used to run model monitoring jobs.

Contents

InstanceCount

The number of ML compute instances to use in the model monitoring job. For distributed processing jobs, specify a value greater than 1. The default value is 1.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: Yes

InstanceType

The ML compute instance type for the processing job.

Type: String

Valid Values: `ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.8xlarge | ml.r5d.12xlarge | ml.r5d.16xlarge | ml.r5d.24xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge |`

ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.p5.4xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge

Required: Yes

VolumeSizeInGB

The size of the ML storage volume, in gigabytes, that you want to provision. You must specify sufficient ML storage for your scenario.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 16384.

Required: Yes

VolumeKmsKeyId

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker AI uses to encrypt data on the storage volume attached to the ML compute instance(s) that run the model monitoring job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringConstraintsResource

Service: Amazon SageMaker Service

The constraints resource for a monitoring job.

Contents

S3Uri

The Amazon S3 URI for the constraints resource.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringCsvDatasetFormat

Service: Amazon SageMaker Service

Represents the CSV dataset format used when running a monitoring job.

Contents

Header

Indicates if the CSV data has a header.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringDatasetFormat

Service: Amazon SageMaker Service

Represents the dataset format used when running a monitoring job.

Contents

Csv

The CSV dataset used in the monitoring job.

Type: MonitoringCsvDatasetFormat object

Required: No

Json

The JSON dataset used in the monitoring job

Type: MonitoringJsonDatasetFormat object

Required: No

Parquet

The Parquet dataset used in the monitoring job

Type: MonitoringParquetDatasetFormat object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringExecutionSummary

Service: Amazon SageMaker Service

Summary of information about the last monitoring job to run.

Contents

CreationTime

The time at which the monitoring job was created.

Type: Timestamp

Required: Yes

LastModifiedTime

A timestamp that indicates the last time the monitoring job was modified.

Type: Timestamp

Required: Yes

MonitoringExecutionStatus

The status of the monitoring job.

Type: String

Valid Values: `Pending | Completed | CompletedWithViolations | InProgress | Failed | Stopping | Stopped`

Required: Yes

MonitoringScheduleName

The name of the monitoring schedule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ScheduledTime

The time the monitoring job was scheduled.

Type: Timestamp

Required: Yes

EndpointName

The name of the endpoint used to run the monitoring job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

FailureReason

Contains the reason a monitoring job failed, if it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

MonitoringJobDefinitionName

The name of the monitoring job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MonitoringType

The type of the monitoring job.

Type: String

Valid Values: `DataQuality | ModelQuality | ModelBias | ModelExplainability`

Required: No

ProcessingJobArn

The Amazon Resource Name (ARN) of the monitoring job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:processing-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringGroundTruthS3Input

Service: Amazon SageMaker Service

The ground truth labels for the dataset used for the monitoring job.

Contents

S3Uri

The address of the Amazon S3 location of the ground truth labels.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringInput

Service: Amazon SageMaker Service

The inputs for a monitoring job.

Contents

BatchTransformInput

Input object for the batch transform job.

Type: BatchTransformInput object

Required: No

EndpointInput

The endpoint for a monitoring job.

Type: EndpointInput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringJobDefinition

Service: Amazon SageMaker Service

Defines the monitoring job.

Contents

MonitoringAppSpecification

Configures the monitoring job to run a specified Docker container image.

Type: MonitoringAppSpecification object

Required: Yes

MonitoringInputs

The array of inputs for the monitoring job. Currently we support monitoring an Amazon SageMaker AI Endpoint.

Type: Array of MonitoringInput objects

Array Members: Fixed number of 1 item.

Required: Yes

MonitoringOutputConfig

The array of outputs from the monitoring job to be uploaded to Amazon S3.

Type: MonitoringOutputConfig object

Required: Yes

MonitoringResources

Identifies the resources, ML compute instances, and ML storage volumes to deploy for a monitoring job. In distributed processing, you specify more than one instance.

Type: MonitoringResources object

Required: Yes

RoleArn

The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker AI can assume to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

BaselineConfig

Baseline configuration used to validate that the data conforms to the specified constraints and statistics

Type: MonitoringBaselineConfig object

Required: No

Environment

Sets the environment variables in the Docker container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[\S\s]*`

Required: No

NetworkConfig

Specifies networking options for an monitoring job.

Type: NetworkConfig object

Required: No

StoppingCondition

Specifies a time limit for how long the monitoring job is allowed to run.

Type: MonitoringStoppingCondition object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringJobDefinitionSummary

Service: Amazon SageMaker Service

Summary information about a monitoring job.

Contents

CreationTime

The time that the monitoring job was created.

Type: Timestamp

Required: Yes

EndpointName

The name of the endpoint that the job monitors.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

MonitoringJobDefinitionArn

The Amazon Resource Name (ARN) of the monitoring job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

MonitoringJobDefinitionName

The name of the monitoring job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringJsonDatasetFormat

Service: Amazon SageMaker Service

Represents the JSON dataset format used when running a monitoring job.

Contents

Line

Indicates if the file should be read as a JSON object per line.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringNetworkConfig

Service: Amazon SageMaker Service

The networking configuration for the monitoring job.

Contents

EnableInterContainerTrafficEncryption

Whether to encrypt all communications between the instances used for the monitoring jobs. Choose `True` to encrypt communications. Encryption provides greater security for distributed jobs, but the processing might take longer.

Type: Boolean

Required: No

EnableNetworkIsolation

Whether to allow inbound and outbound network calls to and from the containers used for the monitoring job.

Type: Boolean

Required: No

VpcConfig

Specifies an Amazon Virtual Private Cloud (VPC) that your SageMaker jobs, hosted models, and compute resources have access to. You can control access to and from your resources by configuring a VPC. For more information, see Give SageMaker Access to Resources in your Amazon VPC.

Type: VpcConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringOutput

Service: Amazon SageMaker Service

The output object for a monitoring job.

Contents

S3Output

The Amazon S3 storage location where the results of a monitoring job are saved.

Type: MonitoringS3Output object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringOutputConfig

Service: Amazon SageMaker Service

The output configuration for monitoring jobs.

Contents

MonitoringOutputs

Monitoring outputs for monitoring jobs. This is where the output of the periodic monitoring jobs is uploaded.

Type: Array of MonitoringOutput objects

Array Members: Fixed number of 1 item.

Required: Yes

KmsKeyId

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker AI uses to encrypt the model artifacts at rest using Amazon S3 server-side encryption.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringParquetDatasetFormat

Service: Amazon SageMaker Service

Represents the Parquet dataset format used when running a monitoring job.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringResources

Service: Amazon SageMaker Service

Identifies the resources to deploy for a monitoring job.

Contents

ClusterConfig

The configuration for the cluster resources used to run the processing job.

Type: MonitoringClusterConfig object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringS3Output

Service: Amazon SageMaker Service

Information about where and how you want to store the results of a monitoring job.

Contents

LocalPath

The local path to the Amazon S3 storage location where Amazon SageMaker AI saves the results of a monitoring job. LocalPath is an absolute path for the output data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

S3Uri

A URI that identifies the Amazon S3 storage location where Amazon SageMaker AI saves the results of a monitoring job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

S3UploadMode

Whether to upload the results of the monitoring job continuously or after the job completes.

Type: String

Valid Values: `Continuous | EndOfJob`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringSchedule

Service: Amazon SageMaker Service

A schedule for a model monitoring job. For information about model monitor, see Amazon SageMaker Model Monitor.

Contents

CreationTime

The time that the monitoring schedule was created.

Type: Timestamp

Required: No

EndpointName

The endpoint that hosts the model being monitored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

FailureReason

If the monitoring schedule failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

LastModifiedTime

The last time the monitoring schedule was changed.

Type: Timestamp

Required: No

LastMonitoringExecutionSummary

Summary of information about the last monitoring job to run.

Type: MonitoringExecutionSummary object

Required: No

MonitoringScheduleArn

The Amazon Resource Name (ARN) of the monitoring schedule.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

MonitoringScheduleConfig

Configures the monitoring schedule and defines the monitoring job.

Type: MonitoringScheduleConfig object

Required: No

MonitoringScheduleName

The name of the monitoring schedule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MonitoringScheduleStatus

The status of the monitoring schedule. This can be one of the following values.

- `PENDING` - The schedule is pending being created.
- `FAILED` - The schedule failed.

- SCHEDULED - The schedule was successfully created.
- STOPPED - The schedule was stopped.

Type: String

Valid Values: `Pending | Failed | Scheduled | Stopped`

Required: No

MonitoringType

The type of the monitoring job definition to schedule.

Type: String

Valid Values: `DataQuality | ModelQuality | ModelBias | ModelExplainability`

Required: No

Tags

A list of the tags associated with the monitoring schedlue. For more information, see Tagging AWS resources in the *AWS General Reference Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringScheduleConfig

Service: Amazon SageMaker Service

Configures the monitoring schedule and defines the monitoring job.

Contents

MonitoringJobDefinition

Defines the monitoring job.

Type: MonitoringJobDefinition object

Required: No

MonitoringJobDefinitionName

The name of the monitoring job definition to schedule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MonitoringType

The type of the monitoring job definition to schedule.

Type: String

Valid Values: `DataQuality | ModelQuality | ModelBias | ModelExplainability`

Required: No

ScheduleConfig

Configures the monitoring schedule.

Type: ScheduleConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringScheduleSummary

Service: Amazon SageMaker Service

Summarizes the monitoring schedule.

Contents

CreationTime

The creation time of the monitoring schedule.

Type: Timestamp

Required: Yes

LastModifiedTime

The last time the monitoring schedule was modified.

Type: Timestamp

Required: Yes

MonitoringScheduleArn

The Amazon Resource Name (ARN) of the monitoring schedule.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: Yes

MonitoringScheduleName

The name of the monitoring schedule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

MonitoringScheduleStatus

The status of the monitoring schedule.

Type: String

Valid Values: `Pending | Failed | Scheduled | Stopped`

Required: Yes

EndpointName

The name of the endpoint using the monitoring schedule.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MonitoringJobDefinitionName

The name of the monitoring job definition that the schedule is for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

MonitoringType

The type of the monitoring job definition that the schedule is for.

Type: String

Valid Values: `DataQuality | ModelQuality | ModelBias | ModelExplainability`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringStatisticsResource

Service: Amazon SageMaker Service

The statistics resource for a monitoring job.

Contents

S3Uri

The Amazon S3 URI for the statistics resource.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonitoringStoppingCondition

Service: Amazon SageMaker Service

A time limit for how long the monitoring job is allowed to run before stopping.

Contents

MaxRuntimeInSeconds

The maximum runtime allowed in seconds.

> **Note**
>
> The `MaxRuntimeInSeconds` cannot exceed the frequency of the job. For data quality and model explainability, this can be up to 3600 seconds for an hourly schedule. For model bias and model quality hourly schedules, this can be up to 1800 seconds.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 86400.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MultiModelConfig

Service: Amazon SageMaker Service

Specifies additional configuration for hosting multi-model endpoints.

Contents

ModelCacheSetting

Whether to cache models for a multi-model endpoint. By default, multi-model endpoints cache models so that a model does not have to be loaded into memory each time it is invoked. Some use cases do not benefit from model caching. For example, if an endpoint hosts a large number of models that are each invoked infrequently, the endpoint might perform better if you disable model caching. To disable model caching, set the value of this parameter to `Disabled`.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NeoVpcConfig

Service: Amazon SageMaker Service

The VpcConfig configuration object that specifies the VPC that you want the compilation jobs to connect to. For more information on controlling access to your Amazon S3 buckets used for compilation job, see Give Amazon SageMaker AI Compilation Jobs Access to Resources in Your Amazon VPC.

Contents

SecurityGroupIds

The VPC security group IDs. IDs have the form of `sg-xxxxxxxx`. Specify the security groups for the VPC that is specified in the `Subnets` field.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: Yes

Subnets

The ID of the subnets in the VPC that you want to connect the compilation job to for accessing the model in Amazon S3.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NestedFilters

Service: Amazon SageMaker Service

A list of nested Filter objects. A resource must satisfy the conditions of all filters to be included in the results returned from the Search API.

For example, to filter on a training job's `InputDataConfig` property with a specific channel name and `S3Uri` prefix, define the following filters:

- `'{Name:"InputDataConfig.ChannelName", "Operator":"Equals", "Value":"train"}',`
- `'{Name:"InputDataConfig.DataSource.S3DataSource.S3Uri", "Operator":"Contains", "Value":"mybucket/catdata"}'`

Contents

Filters

A list of filters. Each filter acts on a property. Filters must contain at least one `Filters` value. For example, a `NestedFilters` call might include a filter on the `PropertyName` parameter of the `InputDataConfig` property: `InputDataConfig.DataSource.S3DataSource.S3Uri`.

Type: Array of Filter objects

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Required: Yes

NestedPropertyName

The name of the property to use in the nested filters. The value must match a listed property name, such as `InputDataConfig`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NetworkConfig

Service: Amazon SageMaker Service

Networking options for a job, such as network traffic encryption between containers, whether to allow inbound and outbound network calls to and from containers, and the VPC subnets and security groups to use for VPC-enabled jobs.

Contents

EnableInterContainerTrafficEncryption

Whether to encrypt all communications between distributed processing jobs. Choose `True` to encrypt communications. Encryption provides greater security for distributed processing jobs, but the processing might take longer.

Type: Boolean

Required: No

EnableNetworkIsolation

Whether to allow inbound and outbound network calls to and from the containers used for the processing job.

Type: Boolean

Required: No

VpcConfig

Specifies an Amazon Virtual Private Cloud (VPC) that your SageMaker jobs, hosted models, and compute resources have access to. You can control access to and from your resources by configuring a VPC. For more information, see Give SageMaker Access to Resources in your Amazon VPC.

Type: VpcConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeAdditionResult

Service: Amazon SageMaker Service

Information about a node that was successfully added to the cluster.

Contents

InstanceGroupName

The name of the instance group to which the node was added.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

NodeLogicalId

A unique identifier assigned to the node that can be used to track its provisioning status through the `DescribeClusterNode` operation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-]*[a-zA-Z0-9]`

Required: Yes

Status

The current status of the node. Possible values include `Pending`, `Running`, `Failed`, `ShuttingDown`, `SystemUpdating`, `DeepHealthCheckInProgress`, and `NotFound`.

Type: String

Valid Values: `Running | Failure | Pending | ShuttingDown | SystemUpdating | DeepHealthCheckInProgress | NotFound`

Required: Yes

AvailabilityZones

The availability zones associated with the successfully added node.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Pattern: `[a-z]{2}-[a-z]+-\d[a-z]`

Required: No

InstanceTypes

The instance types associated with the successfully added node.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5.4xlarge | ml.p6e-gb200.36xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.c5n.large | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.gr6.4xlarge | ml.gr6.8xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.trn2.3xlarge | ml.trn2.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge |`

`ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.i3en.large | ml.i3en.xlarge | ml.i3en.2xlarge | ml.i3en.3xlarge | ml.i3en.6xlarge | ml.i3en.12xlarge | ml.i3en.24xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.r5d.16xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p6-b300.48xlarge`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NotebookInstanceLifecycleConfigSummary

Service: Amazon SageMaker Service

Provides a summary of a notebook instance lifecycle configuration.

Contents

NotebookInstanceLifecycleConfigArn

The Amazon Resource Name (ARN) of the lifecycle configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

NotebookInstanceLifecycleConfigName

The name of the lifecycle configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

CreationTime

A timestamp that tells when the lifecycle configuration was created.

Type: Timestamp

Required: No

LastModifiedTime

A timestamp that tells when the lifecycle configuration was last modified.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NotebookInstanceLifecycleHook

Service: Amazon SageMaker Service

Contains the notebook instance lifecycle configuration script.

Each lifecycle configuration script has a limit of 16384 characters.

The value of the $PATH environment variable that is available to both scripts is `/sbin:bin:/usr/sbin:/usr/bin`.

View Amazon CloudWatch Logs for notebook instance lifecycle configurations in log group `/aws/sagemaker/NotebookInstances` in log stream `[notebook-instance-name]/[LifecycleConfigHook]`.

Lifecycle configuration scripts cannot run for longer than 5 minutes. If a script runs for longer than 5 minutes, it fails and the notebook instance is not created or started.

For information about notebook instance lifestyle configurations, see Step 2.1: (Optional) Customize a Notebook Instance.

Contents

Content

A base64-encoded string that contains a shell script for a notebook instance lifecycle configuration.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 16384.

Pattern: `[\S\s]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

NotebookInstanceSummary

Service: Amazon SageMaker Service

Provides summary information for an SageMaker AI notebook instance.

Contents

NotebookInstanceArn

The Amazon Resource Name (ARN) of the notebook instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

NotebookInstanceName

The name of the notebook instance that you want a summary for.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

AdditionalCodeRepositories

An array of up to three Git repositories associated with the notebook instance. These can be either the names of Git repositories stored as resources in your account, or the URL of Git repositories in AWS CodeCommit or in any other Git repository. These repositories are cloned at the same level as the default repository of your notebook instance. For more information, see Associating Git Repositories with SageMaker AI Notebook Instances.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 3 items.

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `https://([^/]+)/?(.*)$|^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

CreationTime

A timestamp that shows when the notebook instance was created.

Type: Timestamp

Required: No

DefaultCodeRepository

The Git repository associated with the notebook instance as its default code repository. This can be either the name of a Git repository stored as a resource in your account, or the URL of a Git repository in AWS CodeCommit or in any other Git repository. When you open a notebook instance, it opens in the directory that contains this repository. For more information, see Associating Git Repositories with SageMaker AI Notebook Instances.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `https://([^/]+)/?(.*)$|^[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

InstanceType

The type of ML compute instance that the notebook instance is running on.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.8xlarge | ml.m5d.12xlarge | ml.m5d.16xlarge | ml.m5d.24xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p3dn.24xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge |`

ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p6-b200.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.m6id.large | ml.m6id.xlarge | ml.m6id.2xlarge | ml.m6id.4xlarge | ml.m6id.8xlarge | ml.m6id.12xlarge | ml.m6id.16xlarge | ml.m6id.24xlarge | ml.m6id.32xlarge | ml.c6id.large | ml.c6id.xlarge | ml.c6id.2xlarge | ml.c6id.4xlarge | ml.c6id.8xlarge | ml.c6id.12xlarge | ml.c6id.16xlarge | ml.c6id.24xlarge | ml.c6id.32xlarge | ml.r6id.large | ml.r6id.xlarge | ml.r6id.2xlarge | ml.r6id.4xlarge | ml.r6id.8xlarge | ml.r6id.12xlarge | ml.r6id.16xlarge | ml.r6id.24xlarge | ml.r6id.32xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.p5.4xlarge | ml.p5en.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge

Required: No

LastModifiedTime

A timestamp that shows when the notebook instance was last modified.

Type: Timestamp

Required: No

NotebookInstanceLifecycleConfigName

The name of a notebook instance lifecycle configuration associated with this notebook instance.

For information about notebook instance lifestyle configurations, see Step 2.1: (Optional) Customize a Notebook Instance.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

NotebookInstanceStatus

The status of the notebook instance.

Type: String

Valid Values: `Pending | InService | Stopping | Stopped | Failed | Deleting | Updating`

Required: No

Url

The URL that you use to connect to the Jupyter notebook running in your notebook instance.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NotificationConfiguration

Service: Amazon SageMaker Service

Configures Amazon SNS notifications of available or expiring work items for work teams.

Contents

NotificationTopicArn

The ARN for the Amazon SNS topic to which notifications should be published.

Type: String

Pattern: `arn:aws[a-z\-]*:sns:[a-z0-9\-]*:[0-9]{12}:[a-zA-Z0-9_.-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ObjectiveStatusCounters

Service: Amazon SageMaker Service

Specifies the number of training jobs that this hyperparameter tuning job launched, categorized by the status of their objective metric. The objective metric status shows whether the final objective metric for the training job has been evaluated by the tuning job and used in the hyperparameter tuning process.

Contents

Failed

The number of training jobs whose final objective metric was not evaluated and used in the hyperparameter tuning process. This typically occurs when the training job failed or did not emit an objective metric.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

Pending

The number of training jobs that are in progress and pending evaluation of their final objective metric.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

Succeeded

The number of training jobs whose final objective metric was evaluated by the hyperparameter tuning job and used in the hyperparameter tuning process.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OfflineStoreConfig

Service: Amazon SageMaker Service

The configuration of an `OfflineStore`.

Provide an `OfflineStoreConfig` in a request to `CreateFeatureGroup` to create an `OfflineStore`.

To encrypt an `OfflineStore` using at rest data encryption, specify AWS Key Management Service (KMS) key ID, or `KMSKeyId`, in `S3StorageConfig`.

Contents

S3StorageConfig

The Amazon Simple Storage (Amazon S3) location of `OfflineStore`.

Type: S3StorageConfig object

Required: Yes

DataCatalogConfig

The meta data of the Glue table for the `OfflineStore`. If not provided, Feature Store auto-generates the table name, database, and catalog when the `OfflineStore` is created. You can optionally provide this configuration to specify custom values. This applies to both Glue and Apache Iceberg table formats.

Type: DataCatalogConfig object

Required: No

DisableGlueTableCreation

Set to `True` to disable the automatic creation of an AWS Glue table when configuring an `OfflineStore`. If set to `True` and `DataCatalogConfig` is provided, Feature Store associates the provided catalog configuration with the feature group without creating a table. In this case, you are responsible for creating and managing the Glue table. If set to `True` without `DataCatalogConfig`, no Glue table is created or associated with the feature group. The `Iceberg` table format is only supported when this is set to `False`.

If set to `False` and `DataCatalogConfig` is provided, Feature Store creates the table using the specified names. If set to `False` without `DataCatalogConfig`, Feature Store auto-generates

the table name following Athena's naming recommendations. This applies to both Glue and Apache Iceberg table formats.

The default value is `False`.

Type: Boolean

Required: No

TableFormat

Format for the offline store table. Supported formats are Glue (Default) and Apache Iceberg.

Type: String

Valid Values: `Default | Glue | Iceberg`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OfflineStoreStatus

Service: Amazon SageMaker Service

The status of `OfflineStore`.

Contents

Status

An `OfflineStore` status.

Type: String

Valid Values: `Active | Blocked | Disabled`

Required: Yes

BlockedReason

The justification for why the OfflineStoreStatus is Blocked (if applicable).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OidcConfig

Service: Amazon SageMaker Service

Use this parameter to configure your OIDC Identity Provider (IdP).

Contents

AuthorizationEndpoint

The OIDC IdP authorization endpoint used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: Yes

ClientId

The OIDC IdP client ID used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[ -~]+`

Required: Yes

ClientSecret

The OIDC IdP client secret used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[ -~]+`

Required: Yes

Issuer

The OIDC IdP issuer used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: Yes

JwksUri

The OIDC IdP JSON Web Key Set (Jwks) URI used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: Yes

LogoutEndpoint

The OIDC IdP logout endpoint used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: Yes

TokenEndpoint

The OIDC IdP token endpoint used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: Yes

UserInfoEndpoint

The OIDC IdP user information endpoint used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: Yes

AuthenticationRequestExtraParams

A string to string map of identifiers specific to the custom identity provider (IdP) being used.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 10 items.

Key Length Constraints: Minimum length of 0. Maximum length of 512.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 512.

Value Pattern: `.*`

Required: No

Scope

An array of string identifiers used to refer to the specific pieces of user data or claims that the client application wants to access.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[!#-\[\]-~]+( [!#-\[\]-~]+)*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OidcConfigForResponse

Service: Amazon SageMaker Service

Your OIDC IdP workforce configuration.

Contents

AuthenticationRequestExtraParams

A string to string map of identifiers specific to the custom identity provider (IdP) being used.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 10 items.

Key Length Constraints: Minimum length of 0. Maximum length of 512.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 512.

Value Pattern: `.*`

Required: No

AuthorizationEndpoint

The OIDC IdP authorization endpoint used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: No

ClientId

The OIDC IdP client ID used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[ -~]+`

Required: No

Issuer

The OIDC IdP issuer used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: No

JwksUri

The OIDC IdP JSON Web Key Set (Jwks) URI used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: No

LogoutEndpoint

The OIDC IdP logout endpoint used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: No

Scope

An array of string identifiers used to refer to the specific pieces of user data or claims that the client application wants to access.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[!#-\[\]-~]+( [!#-\[\]-~]+)*`

Required: No

TokenEndpoint

The OIDC IdP token endpoint used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: No

UserInfoEndpoint

The OIDC IdP user information endpoint used to configure your private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 500.

Pattern: `https://\S+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OidcMemberDefinition

Service: Amazon SageMaker Service

A list of user groups that exist in your OIDC Identity Provider (IdP). One to ten groups can be used to create a single private work team. When you add a user group to the list of `Groups`, you can add that user group to one or more private work teams. If you add a user group to a private work team, all workers in that user group are added to the work team.

Contents

Groups

A list of comma seperated strings that identifies user groups in your OIDC IdP. Each user group is made up of a group of private workers.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[\p{L}\p{M}\p{S}\p{N}\p{P}]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OnlineStoreConfig

Service: Amazon SageMaker Service

Use this to specify the AWS Key Management Service (KMS) Key ID, or `KMSKeyId`, for at rest data encryption. You can turn `OnlineStore` on or off by specifying the `EnableOnlineStore` flag at General Assembly.

The default value is `False`.

Contents

EnableOnlineStore

Turn `OnlineStore` off by specifying `False` for the `EnableOnlineStore` flag. Turn `OnlineStore` on by specifying `True` for the `EnableOnlineStore` flag.

The default value is `False`.

Type: Boolean

Required: No

SecurityConfig

Use to specify KMS Key ID (`KMSKeyId`) for at-rest encryption of your `OnlineStore`.

Type: OnlineStoreSecurityConfig object

Required: No

StorageType

Option for different tiers of low latency storage for real-time data retrieval.

- `Standard`: A managed low latency data store for feature groups.
- `InMemory`: A managed data store for feature groups that supports very low latency retrieval.

Type: String

Valid Values: `Standard | InMemory`

Required: No

TtlDuration

Time to live duration, where the record is hard deleted after the expiration time is reached; `ExpiresAt` = `EventTime` + `TtlDuration`. For information on HardDelete, see the DeleteRecord API in the Amazon SageMaker API Reference guide.

Type: TtlDuration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OnlineStoreConfigUpdate

Service: Amazon SageMaker Service

Updates the feature group online store configuration.

Contents

TtlDuration

Time to live duration, where the record is hard deleted after the expiration time is reached; `ExpiresAt` = `EventTime` + `TtlDuration`. For information on HardDelete, see the DeleteRecord API in the Amazon SageMaker API Reference guide.

Type: TtlDuration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OnlineStoreSecurityConfig

Service: Amazon SageMaker Service

The security configuration for `OnlineStore`.

Contents

KmsKeyId

The AWS Key Management Service (KMS) key ARN that SageMaker Feature Store uses to encrypt the Amazon S3 objects at rest using Amazon S3 server-side encryption.

The caller (either user or IAM role) of `CreateFeatureGroup` must have below permissions to the `OnlineStore` `KmsKeyId`:

- `"kms:Encrypt"`
- `"kms:Decrypt"`
- `"kms:DescribeKey"`
- `"kms:CreateGrant"`
- `"kms:RetireGrant"`
- `"kms:ReEncryptFrom"`
- `"kms:ReEncryptTo"`
- `"kms:GenerateDataKey"`
- `"kms:ListAliases"`
- `"kms:ListGrants"`
- `"kms:RevokeGrant"`

The caller (either user or IAM role) to all DataPlane operations (`PutRecord`, `GetRecord`, `DeleteRecord`) must have the following permissions to the `KmsKeyId`:

- `"kms:Decrypt"`

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizationConfig

Service: Amazon SageMaker Service

Settings for an optimization technique that you apply with a model optimization job.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

ModelCompilationConfig

Settings for the model compilation technique that's applied by a model optimization job.

Type: ModelCompilationConfig object

Required: No

ModelQuantizationConfig

Settings for the model quantization technique that's applied by a model optimization job.

Type: ModelQuantizationConfig object

Required: No

ModelShardingConfig

Settings for the model sharding technique that's applied by a model optimization job.

Type: ModelShardingConfig object

Required: No

ModelSpeculativeDecodingConfig

Settings for the model speculative decoding technique that's applied by a model optimization job.

Type: ModelSpeculativeDecodingConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizationJobModelSource

Service: Amazon SageMaker Service

The location of the source model to optimize with an optimization job.

Contents

S3

The Amazon S3 location of a source model to optimize with an optimization job.

Type: OptimizationJobModelSourceS3 object

Required: No

SageMakerModel

The name of an existing SageMaker model to optimize with an optimization job.

Type: OptimizationSageMakerModel object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizationJobModelSourceS3

Service: Amazon SageMaker Service

The Amazon S3 location of a source model to optimize with an optimization job.

Contents

ModelAccessConfig

The access configuration settings for the source ML model for an optimization job, where you can accept the model end-user license agreement (EULA).

Type: OptimizationModelAccessConfig object

Required: No

S3Uri

An Amazon S3 URI that locates a source model to optimize with an optimization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizationJobOutputConfig

Service: Amazon SageMaker Service

Details for where to store the optimized model that you create with the optimization job.

Contents

S3OutputLocation

The Amazon S3 URI for where to store the optimized model that you create with an optimization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

KmsKeyId

The Amazon Resource Name (ARN) of a key in AWS KMS. SageMaker uses they key to encrypt the artifacts of the optimized model when SageMaker uploads the model to Amazon S3.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

SageMakerModel

The name of a SageMaker model to use as the output destination for an optimization job.

Type: OptimizationSageMakerModel object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizationJobSummary

Service: Amazon SageMaker Service

Summarizes an optimization job by providing some of its key properties.

Contents

CreationTime

The time when you created the optimization job.

Type: Timestamp

Required: Yes

DeploymentInstanceType

The type of instance that hosts the optimized model that you create with the optimization job.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge`

Required: Yes

OptimizationJobArn

The Amazon Resource Name (ARN) of the optimization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:optimization-job/.*`

Required: Yes

OptimizationJobName

The name that you assigned to the optimization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

OptimizationJobStatus

The current status of the optimization job.

Type: String

Valid Values: `INPROGRESS | COMPLETED | FAILED | STARTING | STOPPING | STOPPED`

Required: Yes

OptimizationTypes

The optimization techniques that are applied by the optimization job.

Type: Array of strings

Required: Yes

LastModifiedTime

The time when the optimization job was last updated.

Type: Timestamp

Required: No

MaxInstanceCount

The maximum number of instances to use for the optimization job.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

OptimizationEndTime

The time when the optimization job finished processing.

Type: Timestamp

Required: No

OptimizationStartTime

The time when the optimization job started.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizationModelAccessConfig

Service: Amazon SageMaker Service

The access configuration settings for the source ML model for an optimization job, where you can accept the model end-user license agreement (EULA).

Contents

AcceptEula

Specifies agreement to the model end-user license agreement (EULA). The `AcceptEula` value must be explicitly defined as `True` in order to accept the EULA that this model requires. You are responsible for reviewing and complying with any applicable license terms and making sure they are acceptable for your use case before downloading or using a model.

Type: Boolean

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizationOutput

Service: Amazon SageMaker Service

Output values produced by an optimization job.

Contents

RecommendedInferenceImage

The image that SageMaker recommends that you use to host the optimized model that you created with an optimization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[\S]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizationSageMakerModel

Service: Amazon SageMaker Service

A SageMaker model to use as the source or destination for an optimization job.

Contents

ModelName

The name of a SageMaker model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizationVpcConfig

Service: Amazon SageMaker Service

A VPC in Amazon VPC that's accessible to an optimized that you create with an optimization job. You can control access to and from your resources by configuring a VPC. For more information, see Give SageMaker Access to Resources in your Amazon VPC.

Contents

SecurityGroupIds

The VPC security group IDs, in the form `sg-xxxxxxxx`. Specify the security groups for the VPC that is specified in the `Subnets` field.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: Yes

Subnets

The ID of the subnets in the VPC to which you want to connect your optimized model.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputConfig

Service: Amazon SageMaker Service

Contains information about the output location for the compiled model and the target device that the model runs on. `TargetDevice` and `TargetPlatform` are mutually exclusive, so you need to choose one between the two to specify your target device or platform. If you cannot find your device you want to use from the `TargetDevice` list, use `TargetPlatform` to describe the platform of your edge device and `CompilerOptions` if there are specific settings that are required or recommended to use for particular TargetPlatform.

Contents

S3OutputLocation

Identifies the S3 bucket where you want Amazon SageMaker AI to store the model artifacts. For example, `s3://bucket-name/key-name-prefix`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

CompilerOptions

Specifies additional parameters for compiler options in JSON format. The compiler options are `TargetPlatform` specific. It is required for NVIDIA accelerators and highly recommended for CPU compilations. For any other cases, it is optional to specify `CompilerOptions`.

- `DTYPE`: Specifies the data type for the input. When compiling for `ml_*` (except for `ml_inf`) instances using PyTorch framework, provide the data type (dtype) of the model's input. `"float32"` is used if `"DTYPE"` is not specified. Options for data type are:
  - float32: Use either `"float"` or `"float32"`.
  - int64: Use either `"int64"` or `"long"`.

  For example, `{"dtype" : "float32"}`.
- `CPU`: Compilation for CPU supports the following compiler options.
  - `mcpu`: CPU micro-architecture. For example, `{'mcpu': 'skylake-avx512'}`
  - `mattr`: CPU flags. For example, `{'mattr': ['+neon', '+vfpv4']}`

- ARM: Details of ARM CPU compilations.
  - NEON: NEON is an implementation of the Advanced SIMD extension used in ARMv7 processors.

    For example, add `{'mattr': ['+neon']}` to the compiler options if compiling for ARM 32-bit platform with the NEON support.
- NVIDIA: Compilation for NVIDIA GPU supports the following compiler options.
  - gpu_code: Specifies the targeted architecture.
  - trt-ver: Specifies the TensorRT versions in x.y.z. format.
  - cuda-ver: Specifies the CUDA version in x.y format.

  For example, `{'gpu-code': 'sm_72', 'trt-ver': '6.0.1', 'cuda-ver': '10.1'}`
- ANDROID: Compilation for the Android OS supports the following compiler options:
  - ANDROID_PLATFORM: Specifies the Android API levels. Available levels range from 21 to 29. For example, `{'ANDROID_PLATFORM': 28}`.
  - mattr: Add `{'mattr': ['+neon']}` to compiler options if compiling for ARM 32-bit platform with NEON support.
- INFERENTIA: Compilation for target ml_inf1 uses compiler options passed in as a JSON string. For example, `"CompilerOptions": "\"--verbose 1 --num-neuroncores 2 -O2\""`.

  For information about supported compiler options, see Neuron Compiler CLI Reference Guide.
- CoreML: Compilation for the CoreML OutputConfig `TargetDevice` supports the following compiler options:
  - class_labels: Specifies the classification labels file name inside input tar.gz file. For example, `{"class_labels": "imagenet_labels_1000.txt"}`. Labels inside the txt file should be separated by newlines.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 1024.

Pattern: `.*`

Required: No

KmsKeyId

The AWS Key Management Service key (AWS KMS) that Amazon SageMaker AI uses to encrypt your output models with Amazon S3 server-side encryption after compilation job. If you don't provide a KMS key ID, Amazon SageMaker AI uses the default KMS key for Amazon S3 for your role's account. For more information, see KMS-Managed Encryption Keys in the *Amazon Simple Storage Service Developer Guide.*

The KmsKeyId can be any of the following formats:

- Key ID: `1234abcd-12ab-34cd-56ef-1234567890ab`
- Key ARN: `arn:aws:kms:us-west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab`
- Alias name: `alias/ExampleAlias`
- Alias name ARN: `arn:aws:kms:us-west-2:111122223333:alias/ExampleAlias`

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

TargetDevice

Identifies the target device or the machine learning instance that you want to run your model on after the compilation has completed. Alternatively, you can specify OS, architecture, and accelerator using TargetPlatform fields. It can be used instead of `TargetPlatform`.

> **Note**
>
> Currently `ml_trn1` is available only in US East (N. Virginia) Region, and `ml_inf2` is available only in US East (Ohio) Region.

Type: String

Valid Values: `lambda | ml_m4 | ml_m5 | ml_m6g | ml_c4 | ml_c5 | ml_c6g | ml_p2 | ml_p3 | ml_g4dn | ml_inf1 | ml_inf2 | ml_trn1 | ml_eia2 | jetson_tx1 | jetson_tx2 | jetson_nano | jetson_xavier | rasp3b | rasp4b`

`| imx8qm | deeplens | rk3399 | rk3288 | aisage | sbe_c | qcs605 | qcs603 | sitara_am57x | amba_cv2 | amba_cv22 | amba_cv25 | x86_win32 | x86_win64 | coreml | jacinto_tda4vm | imx8mplus`

Required: No

TargetPlatform

Contains information about a target platform that you want your model to run on, such as OS, architecture, and accelerators. It is an alternative of `TargetDevice`.

The following examples show how to configure the `TargetPlatform` and `CompilerOptions` JSON strings for popular target platforms:

- Raspberry Pi 3 Model B+

  ```
  "TargetPlatform": {"Os": "LINUX", "Arch": "ARM_EABIHF"},

  "CompilerOptions": {'mattr': ['+neon']}
  ```

- Jetson TX2

  ```
  "TargetPlatform": {"Os": "LINUX", "Arch": "ARM64", "Accelerator": "NVIDIA"},

  "CompilerOptions": {'gpu-code': 'sm_62', 'trt-ver': '6.0.1', 'cuda-ver': '10.0'}
  ```

- EC2 m5.2xlarge instance OS

  ```
  "TargetPlatform": {"Os": "LINUX", "Arch": "X86_64", "Accelerator": "NVIDIA"},

  "CompilerOptions": {'mcpu': 'skylake-avx512'}
  ```

- RK3399

  ```
  "TargetPlatform": {"Os": "LINUX", "Arch": "ARM64", "Accelerator": "MALI"}
  ```

- ARMv7 phone (CPU)

  ```
  "TargetPlatform": {"Os": "ANDROID", "Arch": "ARM_EABI"},

  "CompilerOptions": {'ANDROID_PLATFORM': 25, 'mattr': ['+neon']}
  ```

- ARMv8 phone (CPU)

  ```
  "TargetPlatform": {"Os": "ANDROID", "Arch": "ARM64"},

  "CompilerOptions": {'ANDROID_PLATFORM': 29}
  ```

  Type: TargetPlatform object

  Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputDataConfig

Service: Amazon SageMaker Service

Provides information about how to store model training results (model artifacts).

Contents

S3OutputPath

Identifies the S3 path where you want SageMaker to store the model artifacts. For example, `s3://bucket-name/key-name-prefix`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

CompressionType

The model output compression type. Select `None` to output an uncompressed model, recommended for large model outputs. Defaults to gzip.

Type: String

Valid Values: `GZIP | NONE`

Required: No

KmsKeyId

The AWS Key Management Service (AWS KMS) key that SageMaker uses to encrypt the model artifacts at rest using Amazon S3 server-side encryption. The `KmsKeyId` can be any of the following formats:

- // KMS Key ID

  `"1234abcd-12ab-34cd-56ef-1234567890ab"`

- // Amazon Resource Name (ARN) of a KMS Key

  `"arn:aws:kms:us-west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab"`

- // KMS Key Alias

  `"alias/ExampleAlias"`

- // Amazon Resource Name (ARN) of a KMS Key Alias

  `"arn:aws:kms:us-west-2:111122223333:alias/ExampleAlias"`

If you use a KMS key ID or an alias of your KMS key, the SageMaker execution role must include permissions to call `kms:Encrypt`. If you don't provide a KMS key ID, SageMaker uses the default KMS key for Amazon S3 for your role's account. For more information, see KMS-Managed Encryption Keys in the *Amazon Simple Storage Service Developer Guide*. If the output data is stored in Amazon S3 Express One Zone, it is encrypted with server-side encryption with Amazon S3 managed keys (SSE-S3). KMS key is not supported for Amazon S3 Express One Zone

The KMS key policy must grant permission to the IAM role that you specify in your `CreateTrainingJob`, `CreateTransformJob`, or `CreateHyperParameterTuningJob` requests. For more information, see Using Key Policies in AWS KMS in the *AWS Key Management Service Developer Guide*.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputParameter

Service: Amazon SageMaker Service

An output parameter of a pipeline step.

Contents

Name

The name of the output parameter.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

Value

The value of the output parameter.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OwnershipSettings

Service: Amazon SageMaker Service

The collection of ownership settings for a space.

Contents

OwnerUserProfileName

The user profile who is the owner of the space.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OwnershipSettingsSummary

Service: Amazon SageMaker Service

Specifies summary information about the ownership settings.

Contents

OwnerUserProfileName

The user profile who is the owner of the space.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ParallelismConfiguration

Service: Amazon SageMaker Service

Configuration that controls the parallelism of the pipeline. By default, the parallelism configuration specified applies to all executions of the pipeline unless overridden.

Contents

MaxParallelExecutionSteps

The max number of steps that can be executed in parallel.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Parameter

Service: Amazon SageMaker Service

Assigns a value to a named Pipeline parameter.

Contents

Name

The name of the parameter to assign a value to. This parameter name must match a named parameter in the pipeline definition.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[A-Za-z0-9\-_]*`

Required: Yes

Value

The literal value for the parameter.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ParameterRange

Service: Amazon SageMaker Service

Defines the possible values for categorical, continuous, and integer hyperparameters to be used by an algorithm.

Contents

CategoricalParameterRangeSpecification

A `CategoricalParameterRangeSpecification` object that defines the possible values for a categorical hyperparameter.

Type: CategoricalParameterRangeSpecification object

Required: No

ContinuousParameterRangeSpecification

A `ContinuousParameterRangeSpecification` object that defines the possible values for a continuous hyperparameter.

Type: ContinuousParameterRangeSpecification object

Required: No

IntegerParameterRangeSpecification

A `IntegerParameterRangeSpecification` object that defines the possible values for an integer hyperparameter.

Type: IntegerParameterRangeSpecification object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ParameterRanges

Service: Amazon SageMaker Service

Specifies ranges of integer, continuous, and categorical hyperparameters that a hyperparameter tuning job searches. The hyperparameter tuning job launches training jobs with hyperparameter values within these ranges to find the combination of values that result in the training job with the best performance as measured by the objective metric of the hyperparameter tuning job.

> **Note**
>
> The maximum number of items specified for `Array Members` refers to the maximum number of hyperparameters for each range and also the maximum for the hyperparameter tuning job itself. That is, the sum of the number of hyperparameters for all the ranges can't exceed the maximum number specified.

Contents

AutoParameters

A list containing hyperparameter names and example values to be used by Autotune to determine optimal ranges for your tuning job.

Type: Array of AutoParameter objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Required: No

CategoricalParameterRanges

The array of CategoricalParameterRange objects that specify ranges of categorical hyperparameters that a hyperparameter tuning job searches.

Type: Array of CategoricalParameterRange objects

Array Members: Minimum number of 0 items. Maximum number of 30 items.

Required: No

ContinuousParameterRanges

The array of ContinuousParameterRange objects that specify ranges of continuous hyperparameters that a hyperparameter tuning job searches.

Type: Array of ContinuousParameterRange objects

Array Members: Minimum number of 0 items. Maximum number of 30 items.

Required: No

IntegerParameterRanges

The array of IntegerParameterRange objects that specify ranges of integer hyperparameters that a hyperparameter tuning job searches.

Type: Array of IntegerParameterRange objects

Array Members: Minimum number of 0 items. Maximum number of 30 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Parent

Service: Amazon SageMaker Service

The trial that a trial component is associated with and the experiment the trial is part of. A component might not be associated with a trial. A component can be associated with multiple trials.

Contents

ExperimentName

The name of the experiment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

TrialName

The name of the trial.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ParentHyperParameterTuningJob

Service: Amazon SageMaker Service

A previously completed or stopped hyperparameter tuning job to be used as a starting point for a new hyperparameter tuning job.

Contents

HyperParameterTuningJobName

The name of the hyperparameter tuning job to be used as a starting point for a new hyperparameter tuning job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PartnerAppConfig

Service: Amazon SageMaker Service

Configuration settings for the SageMaker Partner AI App.

Contents

AdminUsers

The list of users that are given admin access to the SageMaker Partner AI App.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(?!\s*$).+`

Required: No

Arguments

This is a map of required inputs for a SageMaker Partner AI App. Based on the application type, the map is populated with a key and value pair that is specific to the user and application.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 5 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `(?!\s*$).+`

Value Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

AssignedGroupPatterns

A list of AWS IAM Identity Center group patterns that can access the SageMaker Partner AI App. Group names support wildcard matching using `*`. An empty list indicates the app will not use Identity Center group features. All groups specified in `RoleGroupAssignments` must match patterns in this list.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\w+=,.@*-]+`

Required: No

RoleGroupAssignments

A map of in-app roles to AWS IAM Identity Center group patterns. Groups assigned to specific roles receive those permissions, while groups in `AssignedGroupPatterns` but not in this map receive default in-app role depending on app type. Group patterns support wildcard matching using `*`. Currently supported by Fiddler version 1.3 and later with roles: `ORG_MEMBER` (default) and `ORG_ADMIN`.

Type: Array of RoleGroupAssignment objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PartnerAppMaintenanceConfig

Service: Amazon SageMaker Service

Maintenance configuration settings for the SageMaker Partner AI App.

Contents

MaintenanceWindowStart

The day and time of the week in Coordinated Universal Time (UTC) 24-hour standard time that weekly maintenance updates are scheduled. This value must take the following format: `3-letter-day:24-h-hour:minute`. For example: `TUE:03:30`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 9.

Pattern: `(Mon|Tue|Wed|Thu|Fri|Sat|Sun):([01]\d|2[0-3]):([0-5]\d)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PartnerAppSummary

Service: Amazon SageMaker Service

A subset of information related to a SageMaker Partner AI App. This information is used as part of the `ListPartnerApps` API response.

Contents

Arn

The ARN of the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:partner-app\/app-[A-Z0-9]{12}`

Required: No

CreationTime

The creation time of the SageMaker Partner AI App.

Type: Timestamp

Required: No

Name

The name of the SageMaker Partner AI App.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]+`

Required: No

Status

The status of the SageMaker Partner AI App.

Type: String

Valid Values: `Creating | Updating | Deleting | Available | Failed | UpdateFailed | Deleted`

Required: No

Type

The type of SageMaker Partner AI App to create. Must be one of the following: `lakera-guard`, `comet`, `deepchecks-llm-evaluation`, or `fiddler`.

Type: String

Valid Values: `lakera-guard | comet | deepchecks-llm-evaluation | fiddler`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PendingDeploymentSummary

Service: Amazon SageMaker Service

The summary of an in-progress deployment when an endpoint is creating or updating with a new endpoint configuration.

Contents

EndpointConfigName

The name of the endpoint configuration used in the deployment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ProductionVariants

An array of PendingProductionVariantSummary objects, one for each model hosted behind this endpoint for the in-progress deployment.

Type: Array of PendingProductionVariantSummary objects

Array Members: Minimum number of 1 item.

Required: No

ShadowProductionVariants

An array of PendingProductionVariantSummary objects, one for each model hosted behind this endpoint in shadow mode with production traffic replicated from the model specified on `ProductionVariants` for the in-progress deployment.

Type: Array of PendingProductionVariantSummary objects

Array Members: Minimum number of 1 item.

Required: No

StartTime

The start time of the deployment.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PendingProductionVariantSummary

Service: Amazon SageMaker Service

The production variant summary for a deployment when an endpoint is creating or updating with the CreateEndpoint or UpdateEndpoint operations. Describes the `VariantStatus` , weight and capacity for a production variant associated with an endpoint.

Contents

VariantName

The name of the variant.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AcceleratorType

This parameter is no longer supported. Elastic Inference (EI) is no longer available.

This parameter was used to specify the size of the EI instance to use for the production variant.

Type: String

Valid Values: `ml.eia1.medium | ml.eia1.large | ml.eia1.xlarge | ml.eia2.medium | ml.eia2.large | ml.eia2.xlarge`

Required: No

CurrentInstanceCount

The number of instances associated with the variant.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

CurrentServerlessConfig

The serverless configuration for the endpoint.

Type: ProductionVariantServerlessConfig object

Required: No

CurrentWeight

The weight associated with the variant.

Type: Float

Valid Range: Minimum value of 0.

Required: No

DeployedImages

An array of `DeployedImage` objects that specify the Amazon EC2 Container Registry paths of the inference images deployed on instances of this `ProductionVariant`.

Type: Array of DeployedImage objects

Required: No

DesiredInstanceCount

The number of instances requested in this deployment, as specified in the endpoint configuration for the endpoint. The value is taken from the request to the CreateEndpointConfig operation.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

DesiredServerlessConfig

The serverless configuration requested for this deployment, as specified in the endpoint configuration for the endpoint.

Type: ProductionVariantServerlessConfig object

Required: No

DesiredWeight

The requested weight for the variant in this deployment, as specified in the endpoint configuration for the endpoint. The value is taken from the request to the CreateEndpointConfig operation.

Type: Float

Valid Range: Minimum value of 0.

Required: No

InstancePools

A list of instance pools for the production variant. Each pool indicates the instance type and the current number of instances of that type.

Type: Array of InstancePoolSummary objects

Array Members: Minimum number of 1 item.

Required: No

InstanceType

The type of instances associated with the variant.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge`

| ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large | ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge | ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large | ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge | ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge | ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge | ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge | ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge | ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge | ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large | ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge | ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge | ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge | ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge | ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge

```
| ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge
| ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge |
ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge
| ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge |
ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge
| ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge |
ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge |
ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium
| ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge |
ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge
| ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge |
ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge
| ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge |
ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge |
ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large |
ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge
| ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge |
ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-
gb200.36xlarge | ml.p5.4xlarge
```

Required: No

ManagedInstanceScaling

Settings that control the range in the number of instances that the endpoint provisions as it scales up or down to accommodate traffic.

Type: ProductionVariantManagedInstanceScaling object

Required: No

RoutingConfig

Settings that control how the endpoint routes incoming traffic to the instances that the endpoint hosts.

Type: ProductionVariantRoutingConfig object

Required: No

VariantStatus

The endpoint variant status which describes the current deployment stage status or operational status.

Type: Array of ProductionVariantStatus objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Phase

Service: Amazon SageMaker Service

Defines the traffic pattern.

Contents

DurationInSeconds

Specifies how long a traffic phase should be. For custom load tests, the value should be between 120 and 3600. This value should not exceed `JobDurationInSeconds`.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

InitialNumberOfUsers

Specifies how many concurrent users to start with. The value should be between 1 and 3.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

SpawnRate

Specified how many new users to spawn in a minute.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

Pipeline

Service: Amazon SageMaker Service

A SageMaker Model Building Pipeline instance.

Contents

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

CreationTime

The creation time of the pipeline.

Type: Timestamp

Required: No

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

LastModifiedTime

The time that the pipeline was last modified.

Type: Timestamp

Required: No

LastRunTime

The time when the pipeline was last run.

Type: Timestamp

Required: No

ParallelismConfiguration

The parallelism configuration applied to the pipeline.

Type: ParallelismConfiguration object

Required: No

PipelineArn

The Amazon Resource Name (ARN) of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

Required: No

PipelineDescription

The description of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

PipelineDisplayName

The display name of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: No

PipelineName

The name of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: No

PipelineStatus

The status of the pipeline.

Type: String

Valid Values: `Active | Deleting`

Required: No

RoleArn

The Amazon Resource Name (ARN) of the role that created the pipeline.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

Tags

A list of tags that apply to the pipeline.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PipelineDefinitionS3Location

Service: Amazon SageMaker Service

The location of the pipeline definition stored in Amazon S3.

Contents

Bucket

Name of the S3 bucket.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 63.

Pattern: `[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9]`

Required: Yes

ObjectKey

The object key (or key name) uniquely identifies the object in an S3 bucket.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `.+`

Required: Yes

VersionId

Version Id of the pipeline definition file. If not specified, Amazon SageMaker will retrieve the latest version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `.+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PipelineExecution

Service: Amazon SageMaker Service

An execution of a pipeline.

Contents

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

CreationTime

The creation time of the pipeline execution.

Type: Timestamp

Required: No

FailureReason

If the execution failed, a message describing why.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1300.

Pattern: `.*`

Required: No

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

LastModifiedTime

The time that the pipeline execution was last modified.

Type: Timestamp

Required: No

ParallelismConfiguration

The parallelism configuration applied to the pipeline execution.

Type: ParallelismConfiguration object

Required: No

PipelineArn

The Amazon Resource Name (ARN) of the pipeline that was executed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

Required: No

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: No

PipelineExecutionDescription

The description of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

PipelineExecutionDisplayName

The display name of the pipeline execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 82.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,81}`

Required: No

PipelineExecutionStatus

The status of the pipeline status.

Type: String

Valid Values: `Executing | Stopping | Stopped | Failed | Succeeded`

Required: No

PipelineExperimentConfig

Specifies the names of the experiment and trial created by a pipeline.

Type: PipelineExperimentConfig object

Required: No

PipelineParameters

Contains a list of pipeline parameters. This list can be empty.

Type: Array of Parameter objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

PipelineVersionDisplayName

The display name of the pipeline version that started this execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 82.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,81}`

Required: No

PipelineVersionId

The ID of the pipeline version that started this execution.

Type: Long

Valid Range: Minimum value of 1.

Required: No

SelectiveExecutionConfig

The selective execution configuration applied to the pipeline run.

Type: SelectiveExecutionConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PipelineExecutionStep

Service: Amazon SageMaker Service

An execution of a step in a pipeline.

Contents

AttemptCount

The current attempt of the execution step. For more information, see Retry Policy for SageMaker Pipelines steps.

Type: Integer

Required: No

CacheHitResult

If this pipeline execution step was cached, details on the cache hit.

Type: CacheHitResult object

Required: No

EndTime

The time that the step stopped executing.

Type: Timestamp

Required: No

FailureReason

The reason why the step failed execution. This is only returned if the step failed its execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

Metadata

Metadata to run the pipeline step.

Type: PipelineExecutionStepMetadata object

Required: No

SelectiveExecutionResult

The ARN from an execution of the current pipeline from which results are reused for this step.

Type: SelectiveExecutionResult object

Required: No

StartTime

The time that the step started executing.

Type: Timestamp

Required: No

StepDescription

The description of the step.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

StepDisplayName

The display name of the step.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

StepName

The name of the step that is executed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[A-Za-z0-9\-_]*`

Required: No

StepStatus

The status of the step execution.

Type: String

Valid Values: `Starting | Executing | Stopping | Stopped | Failed | Succeeded`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PipelineExecutionStepMetadata

Service: Amazon SageMaker Service

Metadata for a step execution.

Contents

AutoMLJob

The Amazon Resource Name (ARN) of the AutoML job that was run by this step.

Type: AutoMLJobStepMetadata object

Required: No

BedrockCustomModel

The metadata of the Amazon Bedrock custom model used in the pipeline execution step.

Type: BedrockCustomModelMetadata object

Required: No

BedrockCustomModelDeployment

The metadata of the Amazon Bedrock custom model deployment used in pipeline execution step.

Type: BedrockCustomModelDeploymentMetadata object

Required: No

BedrockModelImport

The metadata of Amazon Bedrock model import used in pipeline execution step.

Type: BedrockModelImportMetadata object

Required: No

BedrockProvisionedModelThroughput

The metadata of the Amazon Bedrock provisioned model throughput used in the pipeline execution step.

Type: BedrockProvisionedModelThroughputMetadata object

Required: No

Callback

The URL of the Amazon SQS queue used by this step execution, the pipeline generated token, and a list of output parameters.

Type: CallbackStepMetadata object

Required: No

ClarifyCheck

Container for the metadata for a Clarify check step. The configurations and outcomes of the check step execution. This includes:

- The type of the check conducted,
- The Amazon S3 URIs of baseline constraints and statistics files to be used for the drift check.
- The Amazon S3 URIs of newly calculated baseline constraints and statistics.
- The model package group name provided.
- The Amazon S3 URI of the violation report if violations detected.
- The Amazon Resource Name (ARN) of check processing job initiated by the step execution.
- The boolean flags indicating if the drift check is skipped.
- If step property `BaselineUsedForDriftCheck` is set the same as `CalculatedBaseline`.

Type: ClarifyCheckStepMetadata object

Required: No

Condition

The outcome of the condition evaluation that was run by this step execution.

Type: ConditionStepMetadata object

Required: No

EMR

The configurations and outcomes of an Amazon EMR step execution.

Type: EMRStepMetadata object

Required: No

Endpoint

The endpoint that was invoked during this step execution.

Type: EndpointStepMetadata object

Required: No

EndpointConfig

The endpoint configuration used to create an endpoint during this step execution.

Type: EndpointConfigStepMetadata object

Required: No

Fail

The configurations and outcomes of a Fail step execution.

Type: FailStepMetadata object

Required: No

InferenceComponent

The metadata of the inference component used in pipeline execution step.

Type: InferenceComponentMetadata object

Required: No

Job

The metadata for a SageMaker job used in a pipeline execution step.

Type: JobStepMetadata object

Required: No

Lambda

The Amazon Resource Name (ARN) of the Lambda function that was run by this step execution and a list of output parameters.

Type: LambdaStepMetadata object

Required: No

Lineage

The metadata of the lineage used in pipeline execution step.

Type: LineageMetadata object

Required: No

Model

The Amazon Resource Name (ARN) of the model that was created by this step execution.

Type: ModelStepMetadata object

Required: No

ProcessingJob

The Amazon Resource Name (ARN) of the processing job that was run by this step execution.

Type: ProcessingJobStepMetadata object

Required: No

QualityCheck

The configurations and outcomes of the check step execution. This includes:

- The type of the check conducted.
- The Amazon S3 URIs of baseline constraints and statistics files to be used for the drift check.
- The Amazon S3 URIs of newly calculated baseline constraints and statistics.
- The model package group name provided.
- The Amazon S3 URI of the violation report if violations detected.
- The Amazon Resource Name (ARN) of check processing job initiated by the step execution.
- The Boolean flags indicating if the drift check is skipped.
- If step property `BaselineUsedForDriftCheck` is set the same as `CalculatedBaseline`.

Type: QualityCheckStepMetadata object

Required: No

RegisterModel

The Amazon Resource Name (ARN) of the model package that the model was registered to by this step execution.

Type: RegisterModelStepMetadata object

Required: No

TrainingJob

The Amazon Resource Name (ARN) of the training job that was run by this step execution.

Type: TrainingJobStepMetadata object

Required: No

TransformJob

The Amazon Resource Name (ARN) of the transform job that was run by this step execution.

Type: TransformJobStepMetadata object

Required: No

TuningJob

The Amazon Resource Name (ARN) of the tuning job that was run by this step execution.

Type: TuningJobStepMetaData object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PipelineExecutionSummary

Service: Amazon SageMaker Service

A pipeline execution summary.

Contents

PipelineExecutionArn

The Amazon Resource Name (ARN) of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: No

PipelineExecutionDescription

The description of the pipeline execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

PipelineExecutionDisplayName

The display name of the pipeline execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 82.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,81}`

Required: No

PipelineExecutionFailureReason

A message generated by SageMaker Pipelines describing why the pipeline execution failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Required: No

PipelineExecutionStatus

The status of the pipeline execution.

Type: String

Valid Values: `Executing | Stopping | Stopped | Failed | Succeeded`

Required: No

StartTime

The start time of the pipeline execution.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PipelineExperimentConfig

Service: Amazon SageMaker Service

Specifies the names of the experiment and trial created by a pipeline.

Contents

ExperimentName

The name of the experiment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

TrialName

The name of the trial.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PipelineSummary

Service: Amazon SageMaker Service

A summary of a pipeline.

Contents

CreationTime

The creation time of the pipeline.

Type: Timestamp

Required: No

LastExecutionTime

The last time that a pipeline execution began.

Type: Timestamp

Required: No

LastModifiedTime

The time that the pipeline was last modified.

Type: Timestamp

Required: No

PipelineArn

The Amazon Resource Name (ARN) of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

Required: No

PipelineDescription

The description of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

PipelineDisplayName

The display name of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: No

PipelineName

The name of the pipeline.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: No

RoleArn

The Amazon Resource Name (ARN) that the pipeline used to execute.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PipelineVersion

Service: Amazon SageMaker Service

The version of the pipeline.

Contents

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

CreationTime

The creation time of the pipeline version.

Type: Timestamp

Required: No

LastExecutedPipelineExecutionArn

The Amazon Resource Name (ARN) of the most recent pipeline execution created from this pipeline version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: No

LastExecutedPipelineExecutionDisplayName

The display name of the most recent pipeline execution created from this pipeline version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 82.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,81}`

Required: No

LastExecutedPipelineExecutionStatus

The status of the most recent pipeline execution created from this pipeline version.

Type: String

Valid Values: `Executing | Stopping | Stopped | Failed | Succeeded`

Required: No

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

LastModifiedTime

The time when the pipeline version was last modified.

Type: Timestamp

Required: No

PipelineArn

The Amazon Resource Name (ARN) of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

Required: No

PipelineVersionDescription

The description of the pipeline version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

PipelineVersionDisplayName

The display name of the pipeline version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 82.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,81}`

Required: No

PipelineVersionId

The ID of the pipeline version.

Type: Long

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PipelineVersionSummary

Service: Amazon SageMaker Service

The summary of the pipeline version.

Contents

CreationTime

The creation time of the pipeline version.

Type: Timestamp

Required: No

LastExecutionPipelineExecutionArn

The Amazon Resource Name (ARN) of the most recent pipeline execution created from this pipeline version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: No

PipelineArn

The Amazon Resource Name (ARN) of the pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:([0-9]{12}|aws):pipeline/.*`

Required: No

PipelineVersionDescription

The description of the pipeline version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 3072.

Pattern: `.*`

Required: No

PipelineVersionDisplayName

The display name of the pipeline version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 82.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,81}`

Required: No

PipelineVersionId

The ID of the pipeline version.

Type: Long

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PlacementSpecification

Service: Amazon SageMaker Service

Specifies how instances should be placed on a specific UltraServer.

Contents

InstanceCount

The number of ML compute instances required to be placed together on the same UltraServer. Minimum value of 1.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

UltraServerId

The unique identifier of the UltraServer where instances should be placed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PredefinedMetricSpecification

Service: Amazon SageMaker Service

A specification for a predefined metric.

Contents

PredefinedMetricType

The metric type. You can only apply SageMaker metric types to SageMaker endpoints.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PresignedUrlAccessConfig

Service: Amazon SageMaker Service

Configuration for accessing hub content through presigned URLs, including license agreement acceptance and URL validation settings.

Contents

AcceptEula

Indicates acceptance of the End User License Agreement (EULA) for gated models. Set to true to acknowledge acceptance of the license terms required for accessing gated content.

Type: Boolean

Required: No

ExpectedS3Url

The expected S3 URL prefix for validation purposes. This parameter helps ensure consistency between the resolved S3 URIs and the deployment configuration, reducing potential compatibility issues.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PriorityClass

Service: Amazon SageMaker Service

Priority class configuration. When included in `PriorityClasses`, these class configurations define how tasks are queued.

Contents

Name

Name of the priority class.

Type: String

Pattern: `[a-z0-9]([-a-z0-9]*[a-z0-9]){0,39}?`

Required: Yes

Weight

Weight of the priority class. The value is within a range from 0 to 100, where 0 is the default.

A weight of 0 is the lowest priority and 100 is the highest. Weight 0 is the default.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 100.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingClusterConfig

Service: Amazon SageMaker Service

Configuration for the cluster used to run a processing job.

Contents

InstanceCount

The number of ML compute instances to use in the processing job. For distributed processing jobs, specify a value greater than 1. The default value is 1.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: Yes

InstanceType

The ML compute instance type for the processing job.

Type: String

Valid Values: `ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.8xlarge | ml.r5d.12xlarge | ml.r5d.16xlarge | ml.r5d.24xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge |`

```
ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge |
ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge |
ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge
| ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge |
ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge
| ml.m6i.32xlarge | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge |
ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge
| ml.c6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge |
ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge
| ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge
| ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge
| ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large
| ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge |
ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge
| ml.p5.4xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge |
ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge
```

Required: Yes

VolumeSizeInGB

The size of the ML storage volume in gigabytes that you want to provision. You must specify sufficient ML storage for your scenario.

> **Note**
>
> Certain Nitro-based instances include local storage with a fixed total size, dependent on the instance type. When using these instances for processing, Amazon SageMaker mounts the local instance storage instead of Amazon EBS gp2 storage. You can't request a `VolumeSizeInGB` greater than the total size of the local instance storage. For a list of instance types that support local instance storage, including the total size per instance type, see Instance Store Volumes.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 16384.

Required: Yes

VolumeKmsKeyId

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt data on the storage volume attached to the ML compute instance(s) that run the processing job.

> **Note**
>
> Certain Nitro-based instances include local storage, dependent on the instance type. Local storage volumes are encrypted using a hardware module on the instance. You can't request a `VolumeKmsKeyId` when using an instance type with local storage.
> For a list of instance types that support local instance storage, see Instance Store Volumes.
> For more information about local instance storage encryption, see SSD Instance Store Volumes.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingFeatureStoreOutput

Service: Amazon SageMaker Service

Configuration for processing job outputs in Amazon SageMaker Feature Store.

Contents

FeatureGroupName

The name of the Amazon SageMaker FeatureGroup to use as the destination for processing job output. Note that your processing script is responsible for putting records into your Feature Store.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9]([_-]*[a-zA-Z0-9]){0,63}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingInput

Service: Amazon SageMaker Service

The inputs for a processing job. The processing input must specify exactly one of either `S3Input` or `DatasetDefinition` types.

Contents

InputName

The name for the processing job input.

Type: String

Required: Yes

AppManaged

When `True`, input operations such as data download are managed natively by the processing job application. When `False` (default), input operations are managed by Amazon SageMaker.

Type: Boolean

Required: No

DatasetDefinition

Configuration for a Dataset Definition input.

Type: DatasetDefinition object

Required: No

S3Input

Configuration for downloading input data from Amazon S3 into the processing container.

Type: ProcessingS3Input object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingJob

Service: Amazon SageMaker Service

An Amazon SageMaker processing job that is used to analyze data and evaluate models. For more information, see Process Data and Evaluate Models.

Contents

AppSpecification

Configuration to run a processing job in a specified container image.

Type: AppSpecification object

Required: No

AutoMLJobArn

The Amazon Resource Name (ARN) of the AutoML job associated with this processing job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

Required: No

CreationTime

The time the processing job was created.

Type: Timestamp

Required: No

Environment

Sets the environment variables in the Docker container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[\S\s]*`

Required: No

ExitMessage

A string, up to one KB in size, that contains metadata from the processing container when the processing job exits.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\S\s]*`

Required: No

ExperimentConfig

Associates a SageMaker job as a trial component with an experiment and trial. Specified when you call the following APIs:

- CreateProcessingJob
- CreateTrainingJob
- CreateTransformJob

Type: ExperimentConfig object

Required: No

FailureReason

A string, up to one KB in size, that contains the reason a processing job failed, if it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

LastModifiedTime

The time the processing job was last modified.

Type: Timestamp

Required: No

MonitoringScheduleArn

The ARN of a monitoring schedule for an endpoint associated with this processing job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

NetworkConfig

Networking options for a job, such as network traffic encryption between containers, whether to allow inbound and outbound network calls to and from containers, and the VPC subnets and security groups to use for VPC-enabled jobs.

Type: NetworkConfig object

Required: No

ProcessingEndTime

The time that the processing job ended.

Type: Timestamp

Required: No

ProcessingInputs

List of input configurations for the processing job.

Type: Array of ProcessingInput objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

ProcessingJobArn

The ARN of the processing job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:processing-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ProcessingJobName

The name of the processing job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

ProcessingJobStatus

The status of the processing job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: No

ProcessingOutputConfig

Configuration for uploading output from the processing container.

Type: ProcessingOutputConfig object

Required: No

ProcessingResources

Identifies the resources, ML compute instances, and ML storage volumes to deploy for a processing job. In distributed training, you specify more than one instance.

Type: ProcessingResources object

Required: No

ProcessingStartTime

The time that the processing job started.

Type: Timestamp

Required: No

RoleArn

The ARN of the role used to create the processing job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

StoppingCondition

Configures conditions under which the processing job should be stopped, such as how long the processing job has been running. After the condition is met, the processing job is stopped.

Type: ProcessingStoppingCondition object

Required: No

Tags

An array of key-value pairs. For more information, see Using Cost Allocation Tags in the *AWS Billing and Cost Management User Guide*.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TrainingJobArn

The ARN of the training job associated with this processing job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingJobStepMetadata

Service: Amazon SageMaker Service

Metadata for a processing job step.

Contents

Arn

The Amazon Resource Name (ARN) of the processing job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:processing-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingJobSummary

Service: Amazon SageMaker Service

Summary of information about a processing job.

Contents

CreationTime

The time at which the processing job was created.

Type: Timestamp

Required: Yes

ProcessingJobArn

The Amazon Resource Name (ARN) of the processing job..

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:processing-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ProcessingJobName

The name of the processing job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

ProcessingJobStatus

The status of the processing job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: Yes

ExitMessage

An optional string, up to one KB in size, that contains metadata from the processing container when the processing job exits.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\S\s]*`

Required: No

FailureReason

A string, up to one KB in size, that contains the reason a processing job failed, if it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

LastModifiedTime

A timestamp that indicates the last time the processing job was modified.

Type: Timestamp

Required: No

ProcessingEndTime

The time at which the processing job completed.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingOutput

Service: Amazon SageMaker Service

Describes the results of a processing job. The processing output must specify exactly one of either `S3Output` or `FeatureStoreOutput` types.

Contents

OutputName

The name for the processing job output.

Type: String

Required: Yes

AppManaged

When `True`, output operations such as data upload are managed natively by the processing job application. When `False` (default), output operations are managed by Amazon SageMaker.

Type: Boolean

Required: No

FeatureStoreOutput

Configuration for processing job outputs in Amazon SageMaker Feature Store. This processing output type is only supported when `AppManaged` is specified.

Type: ProcessingFeatureStoreOutput object

Required: No

S3Output

Configuration for processing job outputs in Amazon S3.

Type: ProcessingS3Output object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingOutputConfig

Service: Amazon SageMaker Service

Configuration for uploading output from the processing container.

Contents

Outputs

An array of outputs configuring the data to upload from the processing container.

Type: Array of ProcessingOutput objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: Yes

KmsKeyId

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt the processing job output. `KmsKeyId` can be an ID of a KMS key, ARN of a KMS key, or alias of a KMS key. The `KmsKeyId` is applied to all outputs.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingResources

Service: Amazon SageMaker Service

Identifies the resources, ML compute instances, and ML storage volumes to deploy for a processing job. In distributed training, you specify more than one instance.

Contents

ClusterConfig

The configuration for the resources in a cluster used to run the processing job.

Type: ProcessingClusterConfig object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingS3Input

Service: Amazon SageMaker Service

Configuration for downloading input data from Amazon S3 into the processing container.

Contents

S3DataType

Whether you use an `S3Prefix` or a `ManifestFile` for the data type. If you choose `S3Prefix`, `S3Uri` identifies a key name prefix. Amazon SageMaker uses all objects with the specified key name prefix for the processing job. If you choose `ManifestFile`, `S3Uri` identifies an object that is a manifest file containing a list of object keys that you want Amazon SageMaker to use for the processing job.

Type: String

Valid Values: `ManifestFile | S3Prefix`

Required: Yes

S3Uri

The URI of the Amazon S3 prefix Amazon SageMaker downloads data required to run a processing job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

LocalPath

The local path in your container where you want Amazon SageMaker to write input data to. `LocalPath` is an absolute path to the input data and must begin with `/opt/ml/processing/`. `LocalPath` is a required parameter when `AppManaged` is `False` (default).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

S3CompressionType

Whether to GZIP-decompress the data in Amazon S3 as it is streamed into the processing container. `Gzip` can only be used when `Pipe` mode is specified as the `S3InputMode`. In `Pipe` mode, Amazon SageMaker streams input data from the source directly to your container without using the EBS volume.

Type: String

Valid Values: `None | Gzip`

Required: No

S3DataDistributionType

Whether to distribute the data from Amazon S3 to all processing instances with `FullyReplicated`, or whether the data from Amazon S3 is sharded by Amazon S3 key, downloading one shard of data to each processing instance.

Type: String

Valid Values: `FullyReplicated | ShardedByS3Key`

Required: No

S3InputMode

Whether to use `File` or `Pipe` input mode. In File mode, Amazon SageMaker copies the data from the input source onto the local ML storage volume before starting your processing container. This is the most commonly used input mode. In `Pipe` mode, Amazon SageMaker streams input data from the source directly to your processing container into named pipes without using the ML storage volume.

Type: String

Valid Values: `Pipe | File`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingS3Output

Service: Amazon SageMaker Service

Configuration for uploading output data to Amazon S3 from the processing container.

Contents

S3UploadMode

Whether to upload the results of the processing job continuously or after the job completes.

Type: String

Valid Values: `Continuous | EndOfJob`

Required: Yes

S3Uri

A URI that identifies the Amazon S3 bucket where you want Amazon SageMaker to save the results of a processing job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

LocalPath

The local path of a directory where you want Amazon SageMaker to upload its contents to Amazon S3. `LocalPath` is an absolute path to a directory containing output files. This directory will be created by the platform and exist when your container's entrypoint is invoked.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProcessingStoppingCondition

Service: Amazon SageMaker Service

Configures conditions under which the processing job should be stopped, such as how long the processing job has been running. After the condition is met, the processing job is stopped.

Contents

MaxRuntimeInSeconds

Specifies the maximum runtime in seconds.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 777600.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProductionVariant

Service: Amazon SageMaker Service

Identifies a model that you want to host and the resources chosen to deploy for hosting it. If you are deploying multiple models, tell SageMaker how to distribute traffic among the models by specifying variant weights. For more information on production variants, check Production variants.

Contents

VariantName

The name of the production variant.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

AcceleratorType

This parameter is no longer supported. Elastic Inference (EI) is no longer available.

This parameter was used to specify the size of the EI instance to use for the production variant.

Type: String

Valid Values: `ml.eia1.medium | ml.eia1.large | ml.eia1.xlarge | ml.eia2.medium | ml.eia2.large | ml.eia2.xlarge`

Required: No

CapacityReservationConfig

Settings for the capacity reservation for the compute instances that SageMaker AI reserves for an endpoint.

Type: ProductionVariantCapacityReservationConfig object

Required: No

ContainerStartupHealthCheckTimeoutInSeconds

The timeout value, in seconds, for your inference container to pass health check by SageMaker Hosting. For more information about health check, see How Your Container Should Respond to Health Check (Ping) Requests.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 3600.

Required: No

CoreDumpConfig

Specifies configuration for a core dump from the model container when the process crashes.

Type: ProductionVariantCoreDumpConfig object

Required: No

EnableSSMAccess

You can use this parameter to turn on native AWS Systems Manager (SSM) access for a production variant behind an endpoint. By default, SSM access is disabled for all production variants behind an endpoint. You can turn on or turn off SSM access for a production variant behind an existing endpoint by creating a new endpoint configuration and calling `UpdateEndpoint`.

Type: Boolean

Required: No

InferenceAmiVersion

Specifies an option from a collection of preconfigured Amazon Machine Image (AMI) images. Each image is configured by AWS with a set of software and driver versions. AWS optimizes these configurations for different machine learning workloads.

By selecting an AMI version, you can ensure that your inference environment is compatible with specific software requirements, such as CUDA driver versions, Linux kernel versions, or AWS Neuron driver versions.

The AMI version names, and their configurations, are the following:

al2-ami-sagemaker-inference-gpu-2

- Accelerator: GPU

- NVIDIA driver version: 535
- CUDA version: 12.2

al2-ami-sagemaker-inference-gpu-2-1

- Accelerator: GPU
- NVIDIA driver version: 535
- CUDA version: 12.2
- NVIDIA Container Toolkit with disabled CUDA-compat mounting

al2-ami-sagemaker-inference-gpu-3-1

- Accelerator: GPU
- NVIDIA driver version: 550
- CUDA version: 12.4
- NVIDIA Container Toolkit with disabled CUDA-compat mounting

al2023-ami-sagemaker-inference-gpu-4-1

- Accelerator: GPU
- NVIDIA driver version: 580
- CUDA version: 13.0
- NVIDIA Container Toolkit with disabled CUDA-compat mounting

al2-ami-sagemaker-inference-neuron-2

- Accelerator: Inferentia2 and Trainium
- Neuron driver version: 2.19

Type: String

Valid Values: `al2-ami-sagemaker-inference-gpu-2 | al2-ami-sagemaker-inference-gpu-2-1 | al2-ami-sagemaker-inference-gpu-3-1 | al2-ami-sagemaker-inference-neuron-2 | al2023-ami-sagemaker-inference-gpu-4-1`

Required: No

InitialInstanceCount

Number of instances to launch initially.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

InitialVariantWeight

Determines initial traffic distribution among all of the models that you specify in the endpoint configuration. The traffic to a production variant is determined by the ratio of the `VariantWeight` to the sum of all `VariantWeight` values across all ProductionVariants. If unspecified, it defaults to 1.0.

Type: Float

Valid Range: Minimum value of 0.

Required: No

InstancePools

A list of instance pools for the production variant. Each instance pool specifies an instance type and its priority for provisioning. Use instance pools to configure heterogeneous endpoints that deploy models across multiple instance types.

Type: Array of InstancePool objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

InstanceType

The ML compute instance type.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge |`

ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large | ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge | ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large | ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge | ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge | ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge | ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge | ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge | ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge | ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large | ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge | ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge | ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge | ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge |

```
ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge
| ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge |
ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge
| ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge
| ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge |
ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge
| ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge |
ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge
| ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge |
ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge |
ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium
| ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge |
ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge
| ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge |
ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge
| ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge |
ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge |
ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large |
ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge
| ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge |
ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-
gb200.36xlarge | ml.p5.4xlarge
```

Required: No

ManagedInstanceScaling

Settings that control the range in the number of instances that the endpoint provisions as it scales up or down to accommodate traffic.

Type: ProductionVariantManagedInstanceScaling object

Required: No

ModelDataDownloadTimeoutInSeconds

The timeout value, in seconds, to download and extract the model that you want to host from Amazon S3 to the individual inference instance associated with this production variant.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 3600.

Required: No

ModelName

The name of the model that you want to host. This is the name that you specified when creating the model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

RoutingConfig

Settings that control how the endpoint routes incoming traffic to the instances that the endpoint hosts.

Type: ProductionVariantRoutingConfig object

Required: No

ServerlessConfig

The serverless configuration for an endpoint. Specifies a serverless endpoint configuration instead of an instance-based endpoint configuration.

Type: ProductionVariantServerlessConfig object

Required: No

VariantInstanceProvisionTimeoutInSeconds

The timeout value, in seconds, for provisioning instances for the production variant. When SageMaker encounters an insufficient capacity error while provisioning instances, it retries with the next instance pool (if configured) or waits until the timeout expires. This timeout applies only to capacity provisioning and does not include the time for model download or container startup.

Valid values: 300 to 3600.

Type: Integer

Valid Range: Minimum value of 300. Maximum value of 3600.

Required: No

VolumeSizeInGB

The size, in GB, of the ML storage volume attached to individual inference instance associated with the production variant. Currently only Amazon EBS gp2 storage volumes are supported.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 512.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProductionVariantCapacityReservationConfig

Service: Amazon SageMaker Service

Settings for the capacity reservation for the compute instances that SageMaker AI reserves for an endpoint.

Contents

CapacityReservationPreference

Options that you can choose for the capacity reservation. SageMaker AI supports the following options:

capacity-reservations-only

SageMaker AI launches instances only into an ML capacity reservation. If no capacity is available, the instances fail to launch.

Type: String

Valid Values: `capacity-reservations-only`

Required: No

MlReservationArn

The Amazon Resource Name (ARN) that uniquely identifies the ML capacity reservation that SageMaker AI applies when it deploys the endpoint.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 258.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z0-9\-]{1,14}/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProductionVariantCapacityReservationSummary

Service: Amazon SageMaker Service

Details about an ML capacity reservation.

Contents

AvailableInstanceCount

The number of instances that are currently available in the ML capacity reservation.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

CapacityReservationPreference

The option that you chose for the capacity reservation. SageMaker AI supports the following options:

capacity-reservations-only

SageMaker AI launches instances only into an ML capacity reservation. If no capacity is available, the instances fail to launch.

Type: String

Valid Values: `capacity-reservations-only`

Required: No

Ec2CapacityReservations

The EC2 capacity reservations that are shared to this ML capacity reservation, if any.

Type: Array of Ec2CapacityReservation objects

Required: No

MlReservationArn

The Amazon Resource Name (ARN) that uniquely identifies the ML capacity reservation that SageMaker AI applies when it deploys the endpoint.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 258.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z0-9\-]{1,14}/.*`

Required: No

TotalInstanceCount

The number of instances that you allocated to the ML capacity reservation.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

UsedByCurrentEndpoint

The number of instances from the ML capacity reservation that are being used by the endpoint.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProductionVariantCoreDumpConfig

Service: Amazon SageMaker Service

Specifies configuration for a core dump from the model container when the process crashes.

Contents

DestinationS3Uri

The Amazon S3 bucket to send the core dump to.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `(https|s3)://([^/])/?(.*)`

Required: Yes

KmsKeyId

The AWS Key Management Service (AWS KMS) key that SageMaker uses to encrypt the core dump data at rest using Amazon S3 server-side encryption. The `KmsKeyId` can be any of the following formats:

- // KMS Key ID

  `"1234abcd-12ab-34cd-56ef-1234567890ab"`
- // Amazon Resource Name (ARN) of a KMS Key

  `"arn:aws:kms:us-west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab"`
- // KMS Key Alias

  `"alias/ExampleAlias"`
- // Amazon Resource Name (ARN) of a KMS Key Alias

  `"arn:aws:kms:us-west-2:111122223333:alias/ExampleAlias"`

If you use a KMS key ID or an alias of your KMS key, the SageMaker execution role must include permissions to call `kms:Encrypt`. If you don't provide a KMS key ID, SageMaker uses the default KMS key for Amazon S3 for your role's account. SageMaker uses server-

side encryption with KMS-managed keys for `OutputDataConfig`. If you use a bucket policy with an `s3:PutObject` permission that only allows objects with server-side encryption, set the condition key of `s3:x-amz-server-side-encryption` to `"aws:kms"`. For more information, see KMS-Managed Encryption Keys in the *Amazon Simple Storage Service Developer Guide.*

The KMS key policy must grant permission to the IAM role that you specify in your `CreateEndpoint` and `UpdateEndpoint` requests. For more information, see Using Key Policies in AWS KMS in the *AWS Key Management Service Developer Guide*.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProductionVariantManagedInstanceScaling

Service: Amazon SageMaker Service

Settings that control the range in the number of instances that the endpoint provisions as it scales up or down to accommodate traffic.

Contents

MaxInstanceCount

The maximum number of instances that the endpoint can provision when it scales up to accommodate an increase in traffic.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

MinInstanceCount

The minimum number of instances that the endpoint must retain when it scales down to accommodate a decrease in traffic.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

ScaleInPolicy

Configures the scale-in behavior for managed instance scaling.

Type: ProductionVariantManagedInstanceScalingScaleInPolicy object

Required: No

Status

Indicates whether managed instance scaling is enabled.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProductionVariantManagedInstanceScalingScaleInPolicy

Service: Amazon SageMaker Service

Configures the scale-in behavior for managed instance scaling.

Contents

Strategy

The strategy for scaling in instances.

IDLE_RELEASE

Releases instances that have no hosted inference component copies.

CONSOLIDATION

Consolidates inference component copies onto fewer instances to release more instances. Consolidation honors the scheduling configuration of each inference component. For example, if an inference component specifies Availability Zone balance, consolidation only proceeds when the resulting distribution does not increase the imbalance.

Type: String

Valid Values: `IDLE_RELEASE | CONSOLIDATION`

Required: Yes

CooldownInMinutes

The cooldown period, in minutes, after the last endpoint operation before the endpoint evaluates consolidation scale-in opportunities.

Default value: `20`.

Type: Integer

Valid Range: Minimum value of 5. Maximum value of 1440.

Required: No

MaximumStepSize

The maximum number of instances that the endpoint can terminate at a time during a consolidation scale-in operation.

Default value: 1.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProductionVariantRoutingConfig

Service: Amazon SageMaker Service

Settings that control how the endpoint routes incoming traffic to the instances that the endpoint hosts.

Contents

RoutingStrategy

Sets how the endpoint routes incoming traffic:

- `LEAST_OUTSTANDING_REQUESTS`: The endpoint routes requests to the specific instances that have more capacity to process them.
- `RANDOM`: The endpoint routes each request to a randomly chosen instance.

Type: String

Valid Values: `LEAST_OUTSTANDING_REQUESTS | RANDOM`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProductionVariantServerlessConfig

Service: Amazon SageMaker Service

Specifies the serverless configuration for an endpoint variant.

Contents

MaxConcurrency

The maximum number of concurrent invocations your serverless endpoint can process.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 200.

Required: Yes

MemorySizeInMB

The memory size of your serverless endpoint. Valid values are in 1 GB increments: 1024 MB, 2048 MB, 3072 MB, 4096 MB, 5120 MB, or 6144 MB.

Type: Integer

Valid Range: Minimum value of 1024. Maximum value of 6144.

Required: Yes

ProvisionedConcurrency

The amount of provisioned concurrency to allocate for the serverless endpoint. Should be less than or equal to `MaxConcurrency`.

> **Note**
>
> This field is not supported for serverless endpoint recommendations for Inference Recommender jobs. For more information about creating an Inference Recommender job, see CreateInferenceRecommendationsJobs.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProductionVariantServerlessUpdateConfig

Service: Amazon SageMaker Service

Specifies the serverless update concurrency configuration for an endpoint variant.

Contents

MaxConcurrency

The updated maximum number of concurrent invocations your serverless endpoint can process.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 200.

Required: No

ProvisionedConcurrency

The updated amount of provisioned concurrency to allocate for the serverless endpoint. Should be less than or equal to `MaxConcurrency`.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProductionVariantStatus

Service: Amazon SageMaker Service

Describes the status of the production variant.

Contents

Status

The endpoint variant status which describes the current deployment stage status or operational status.

- `Creating`: Creating inference resources for the production variant.
- `Deleting`: Terminating inference resources for the production variant.
- `Updating`: Updating capacity for the production variant.
- `ActivatingTraffic`: Turning on traffic for the production variant.
- `Baking`: Waiting period to monitor the CloudWatch alarms in the automatic rollback configuration.

Type: String

Valid Values: `Creating | Updating | Deleting | ActivatingTraffic | Baking`

Required: Yes

StartTime

The start time of the current status change.

Type: Timestamp

Required: No

StatusMessage

A message that describes the status of the production variant.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProductionVariantSummary

Service: Amazon SageMaker Service

Describes weight and capacities for a production variant associated with an endpoint. If you sent a request to the `UpdateEndpointWeightsAndCapacities` API and the endpoint status is `Updating`, you get different desired and current values.

Contents

VariantName

The name of the variant.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

CapacityReservationConfig

Settings for the capacity reservation for the compute instances that SageMaker AI reserves for an endpoint.

Type: ProductionVariantCapacityReservationSummary object

Required: No

CurrentInstanceCount

The number of instances associated with the variant.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

CurrentServerlessConfig

The serverless configuration for the endpoint.

Type: ProductionVariantServerlessConfig object

Required: No

CurrentWeight

The weight associated with the variant.

Type: Float

Valid Range: Minimum value of 0.

Required: No

DeployedImages

An array of `DeployedImage` objects that specify the Amazon EC2 Container Registry paths of the inference images deployed on instances of this `ProductionVariant`.

Type: Array of DeployedImage objects

Required: No

DesiredInstanceCount

The number of instances requested in the `UpdateEndpointWeightsAndCapacities` request.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

DesiredServerlessConfig

The serverless configuration requested for the endpoint update.

Type: ProductionVariantServerlessConfig object

Required: No

DesiredWeight

The requested weight, as specified in the `UpdateEndpointWeightsAndCapacities` request.

Type: Float

Valid Range: Minimum value of 0.

Required: No

InstancePools

A list of instance pools for the production variant. Each pool indicates the instance type and the current number of instances of that type.

Type: Array of InstancePoolSummary objects

Array Members: Minimum number of 1 item.

Required: No

ManagedInstanceScaling

Settings that control the range in the number of instances that the endpoint provisions as it scales up or down to accommodate traffic.

Type: ProductionVariantManagedInstanceScaling object

Required: No

RoutingConfig

Settings that control how the endpoint routes incoming traffic to the instances that the endpoint hosts.

Type: ProductionVariantRoutingConfig object

Required: No

VariantStatus

The endpoint variant status which describes the current deployment stage status or operational status.

Type: Array of ProductionVariantStatus objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProfilerConfig

Service: Amazon SageMaker Service

Configuration information for Amazon SageMaker Debugger system monitoring, framework profiling, and storage paths.

Contents

DisableProfiler

Configuration to turn off Amazon SageMaker Debugger's system monitoring and profiling functionality. To turn it off, set to `True`.

Type: Boolean

Required: No

ProfilingIntervalInMilliseconds

A time interval for capturing system metrics in milliseconds. Available values are 100, 200, 500, 1000 (1 second), 5000 (5 seconds), and 60000 (1 minute) milliseconds. The default value is 500 milliseconds.

Type: Long

Required: No

ProfilingParameters

Configuration information for capturing framework metrics. Available key strings for different profiling options are `DetailedProfilingConfig`, `PythonProfilingConfig`, and `DataLoaderProfilingConfig`. The following codes are configuration structures for the `ProfilingParameters` parameter. To learn more about how to configure the `ProfilingParameters` parameter, see Use the SageMaker and Debugger Configuration API Operations to Create, Update, and Debug Your Training Job.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 20 items.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `.*`

Required: No

S3OutputPath

Path to Amazon S3 storage location for system and framework metrics.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProfilerConfigForUpdate

Service: Amazon SageMaker Service

Configuration information for updating the Amazon SageMaker Debugger profile parameters, system and framework metrics configurations, and storage paths.

Contents

DisableProfiler

To turn off Amazon SageMaker Debugger monitoring and profiling while a training job is in progress, set to `True`.

Type: Boolean

Required: No

ProfilingIntervalInMilliseconds

A time interval for capturing system metrics in milliseconds. Available values are 100, 200, 500, 1000 (1 second), 5000 (5 seconds), and 60000 (1 minute) milliseconds. The default value is 500 milliseconds.

Type: Long

Required: No

ProfilingParameters

Configuration information for capturing framework metrics. Available key strings for different profiling options are `DetailedProfilingConfig`, `PythonProfilingConfig`, and `DataLoaderProfilingConfig`. The following codes are configuration structures for the `ProfilingParameters` parameter. To learn more about how to configure the `ProfilingParameters` parameter, see Use the SageMaker and Debugger Configuration API Operations to Create, Update, and Debug Your Training Job.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 20 items.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `.*`

Required: No

S3OutputPath

Path to Amazon S3 storage location for system and framework metrics.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProfilerRuleConfiguration

Service: Amazon SageMaker Service

Configuration information for profiling rules.

Contents

RuleConfigurationName

The name of the rule configuration. It must be unique relative to other rule configuration names.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*`

Required: Yes

RuleEvaluatorImage

The Amazon Elastic Container Registry Image for the managed rule evaluation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `.*`

Required: Yes

InstanceType

The instance type to deploy a custom rule for profiling a training job.

Type: String

Valid Values: `ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge`

| ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.8xlarge | ml.r5d.12xlarge | ml.r5d.16xlarge | ml.r5d.24xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.p5.4xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge

Required: No

LocalPath

Path to local storage location for output of rules. Defaults to `/opt/ml/processing/output/rule/`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4096.

Pattern: `.*`

Required: No

RuleParameters

Runtime configuration for rule container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `.*`

Required: No

S3OutputPath

Path to Amazon S3 storage location for rules.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

VolumeSizeInGB

The size, in GB, of the ML storage volume attached to the processing instance.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProfilerRuleEvaluationStatus

Service: Amazon SageMaker Service

Information about the status of the rule evaluation.

Contents

LastModifiedTime

Timestamp when the rule evaluation status was last modified.

Type: Timestamp

Required: No

RuleConfigurationName

The name of the rule configuration.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*`

Required: No

RuleEvaluationJobArn

The Amazon Resource Name (ARN) of the rule evaluation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:processing-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

RuleEvaluationStatus

Status of the rule evaluation.

Type: String

Valid Values: `InProgress | NoIssuesFound | IssuesFound | Error | Stopping | Stopped`

Required: No

StatusDetails

Details from the rule evaluation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Project

Service: Amazon SageMaker Service

The properties of a project as returned by the Search API.

Contents

CreatedBy

Who created the project.

Type: UserContext object

Required: No

CreationTime

A timestamp specifying when the project was created.

Type: Timestamp

Required: No

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

LastModifiedTime

A timestamp container for when the project was last modified.

Type: Timestamp

Required: No

ProjectArn

The Amazon Resource Name (ARN) of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:project/[\S]{1,2048}`

Required: No

ProjectDescription

The description of the project.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

ProjectId

The ID of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

ProjectName

The name of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: No

ProjectStatus

The status of the project.

Type: String

Valid Values: `Pending | CreateInProgress | CreateCompleted | CreateFailed | DeleteInProgress | DeleteFailed | DeleteCompleted | UpdateInProgress | UpdateCompleted | UpdateFailed`

Required: No

ServiceCatalogProvisionedProductDetails

Details of a provisioned service catalog product. For information about service catalog, see What is AWS Service Catalog.

Type: ServiceCatalogProvisionedProductDetails object

Required: No

ServiceCatalogProvisioningDetails

Details that you specify to provision a service catalog product. For information about service catalog, see What is AWS Service Catalog.

Type: ServiceCatalogProvisioningDetails object

Required: No

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TemplateProviderDetails

An array of template providers associated with the project.

Type: Array of TemplateProviderDetail objects

Array Members: Fixed number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProjectSummary

Service: Amazon SageMaker Service

Information about a project.

Contents

CreationTime

The time that the project was created.

Type: Timestamp

Required: Yes

ProjectArn

The Amazon Resource Name (ARN) of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:project/[\S]{1,2048}`

Required: Yes

ProjectId

The ID of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

ProjectName

The name of the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,31}`

Required: Yes

ProjectStatus

The status of the project.

Type: String

Valid Values: `Pending | CreateInProgress | CreateCompleted | CreateFailed | DeleteInProgress | DeleteFailed | DeleteCompleted | UpdateInProgress | UpdateCompleted | UpdateFailed`

Required: Yes

ProjectDescription

The description of the project.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\p{L}\p{M}\p{Z}\p{S}\p{N}\p{P}]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PropertyNameQuery

Service: Amazon SageMaker Service

Part of the `SuggestionQuery` type. Specifies a hint for retrieving property names that begin with the specified text.

Contents

PropertyNameHint

Text that begins a property's name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 100.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PropertyNameSuggestion

Service: Amazon SageMaker Service

A property name returned from a `GetSearchSuggestions` call that specifies a value in the `PropertyNameQuery` field.

Contents

PropertyName

A suggested property name based on what you entered in the search textbox in the SageMaker console.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProvisioningParameter

Service: Amazon SageMaker Service

A key value pair used when you provision a project as a service catalog product. For information, see What is AWS Service Catalog.

Contents

Key

The key that identifies a provisioning parameter.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1000.

Pattern: `.*`

Required: No

Value

The value of the provisioning parameter.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4096.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PublicWorkforceTaskPrice

Service: Amazon SageMaker Service

Defines the amount of money paid to an Amazon Mechanical Turk worker for each task performed.

Use one of the following prices for bounding box tasks. Prices are in US dollars and should be based on the complexity of the task; the longer it takes in your initial testing, the more you should offer.

- 0.036
- 0.048
- 0.060
- 0.072
- 0.120
- 0.240
- 0.360
- 0.480
- 0.600
- 0.720
- 0.840
- 0.960
- 1.080
- 1.200

Use one of the following prices for image classification, text classification, and custom tasks. Prices are in US dollars.

- 0.012
- 0.024
- 0.036
- 0.048
- 0.060
- 0.072

- 0.120
- 0.240
- 0.360
- 0.480
- 0.600
- 0.720
- 0.840
- 0.960
- 1.080
- 1.200

Use one of the following prices for semantic segmentation tasks. Prices are in US dollars.

- 0.840
- 0.960
- 1.080
- 1.200

Use one of the following prices for Textract AnalyzeDocument Important Form Key Amazon Augmented AI review tasks. Prices are in US dollars.

- 2.400
- 2.280
- 2.160
- 2.040
- 1.920
- 1.800
- 1.680
- 1.560
- 1.440
- 1.320
- 1.200

- 1.080
- 0.960
- 0.840
- 0.720
- 0.600
- 0.480
- 0.360
- 0.240
- 0.120
- 0.072
- 0.060
- 0.048
- 0.036
- 0.024
- 0.012

Use one of the following prices for Rekognition DetectModerationLabels Amazon Augmented AI review tasks. Prices are in US dollars.

- 1.200
- 1.080
- 0.960
- 0.840
- 0.720
- 0.600
- 0.480
- 0.360
- 0.240
- 0.120
- 0.072
- 0.060

- 0.048
- 0.036
- 0.024
- 0.012

Use one of the following prices for Amazon Augmented AI custom human review tasks. Prices are in US dollars.

- 1.200
- 1.080
- 0.960
- 0.840
- 0.720
- 0.600
- 0.480
- 0.360
- 0.240
- 0.120
- 0.072
- 0.060
- 0.048
- 0.036
- 0.024
- 0.012

Contents

AmountInUsd

Defines the amount of money paid to an Amazon Mechanical Turk worker in United States dollars.

Type: USD object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QualityCheckStepMetadata

Service: Amazon SageMaker Service

Container for the metadata for a Quality check step. For more information, see the topic on QualityCheck step in the *Amazon SageMaker Developer Guide*.

Contents

BaselineUsedForDriftCheckConstraints

The Amazon S3 URI of the baseline constraints file used for the drift check.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

BaselineUsedForDriftCheckStatistics

The Amazon S3 URI of the baseline statistics file used for the drift check.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

CalculatedBaselineConstraints

The Amazon S3 URI of the newly calculated baseline constraints file.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

CalculatedBaselineStatistics

The Amazon S3 URI of the newly calculated baseline statistics file.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

CheckJobArn

The Amazon Resource Name (ARN) of the Quality check processing job that was run by this step execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

CheckType

The type of the Quality check step.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

ModelPackageGroupName

The model package group name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

RegisterNewBaseline

This flag indicates if a newly calculated baseline can be accessed through step properties `BaselineUsedForDriftCheckConstraints` and `BaselineUsedForDriftCheckStatistics`. If it is set to `False`, the previous baseline of the configured check type must also be available. These can be accessed through the `BaselineUsedForDriftCheckConstraints` and `BaselineUsedForDriftCheckStatistics` properties.

Type: Boolean

Required: No

SkipCheck

This flag indicates if the drift check against the previous baseline will be skipped or not. If it is set to `False`, the previous baseline of the configured check type must be available.

Type: Boolean

Required: No

ViolationReport

The Amazon S3 URI of violation report if violations are detected.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QueryFilters

Service: Amazon SageMaker Service

A set of filters to narrow the set of lineage entities connected to the `StartArn`(s) returned by the `QueryLineage` API action.

Contents

CreatedAfter

Filter the lineage entities connected to the `StartArn`(s) after the create date.

Type: Timestamp

Required: No

CreatedBefore

Filter the lineage entities connected to the `StartArn`(s) by created date.

Type: Timestamp

Required: No

LineageTypes

Filter the lineage entities connected to the `StartArn`(s) by the type of the lineage entity.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Valid Values: `TrialComponent | Artifact | Context | Action`

Required: No

ModifiedAfter

Filter the lineage entities connected to the `StartArn`(s) after the last modified date.

Type: Timestamp

Required: No

ModifiedBefore

Filter the lineage entities connected to the `StartArn`(s) before the last modified date.

Type: Timestamp

Required: No

Properties

Filter the lineage entities connected to the `StartArn`(s) by a set if property key value pairs. If multiple pairs are provided, an entity is included in the results if it matches any of the provided pairs.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 5 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

Types

Filter the lineage entities connected to the `StartArn` by type. For example: `DataSet`, `Model`, `Endpoint`, or `ModelDeployment`.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 40.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

RealTimeInferenceConfig

Service: Amazon SageMaker Service

The infrastructure configuration for deploying the model to a real-time inference endpoint.

Contents

InstanceCount

The number of instances of the type specified by `InstanceType`.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

InstanceType

The instance type the model is deployed to.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge |`

ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large | ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge | ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large | ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge | ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge | ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge | ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge | ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge | ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge | ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large | ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge | ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge | ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge | ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge | ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge |

`ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium | ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge | ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge | ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge | ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge | ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge | ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge | ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large | ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge | ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge | ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RealTimeInferenceRecommendation

Service: Amazon SageMaker Service

The recommended configuration to use for Real-Time Inference.

Contents

InstanceType

The recommended instance type for Real-Time Inference.

Type: String

Valid Values: `ml.t2.medium | ml.t2.large | ml.t2.xlarge | ml.t2.2xlarge | ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.12xlarge | ml.m5d.24xlarge | ml.c4.large | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5d.large | ml.c5d.xlarge | ml.c5d.2xlarge | ml.c5d.4xlarge | ml.c5d.9xlarge | ml.c5d.18xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.12xlarge | ml.r5.24xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.12xlarge | ml.r5d.24xlarge | ml.inf1.xlarge | ml.inf1.2xlarge | ml.inf1.6xlarge | ml.inf1.24xlarge | ml.dl1.24xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge`

| ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.r8g.medium | ml.r8g.large | ml.r8g.xlarge | ml.r8g.2xlarge | ml.r8g.4xlarge | ml.r8g.8xlarge | ml.r8g.12xlarge | ml.r8g.16xlarge | ml.r8g.24xlarge | ml.r8g.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge | ml.p4d.24xlarge | ml.c7g.large | ml.c7g.xlarge | ml.c7g.2xlarge | ml.c7g.4xlarge | ml.c7g.8xlarge | ml.c7g.12xlarge | ml.c7g.16xlarge | ml.m6g.large | ml.m6g.xlarge | ml.m6g.2xlarge | ml.m6g.4xlarge | ml.m6g.8xlarge | ml.m6g.12xlarge | ml.m6g.16xlarge | ml.m6gd.large | ml.m6gd.xlarge | ml.m6gd.2xlarge | ml.m6gd.4xlarge | ml.m6gd.8xlarge | ml.m6gd.12xlarge | ml.m6gd.16xlarge | ml.c6g.large | ml.c6g.xlarge | ml.c6g.2xlarge | ml.c6g.4xlarge | ml.c6g.8xlarge | ml.c6g.12xlarge | ml.c6g.16xlarge | ml.c6gd.large | ml.c6gd.xlarge | ml.c6gd.2xlarge | ml.c6gd.4xlarge | ml.c6gd.8xlarge | ml.c6gd.12xlarge | ml.c6gd.16xlarge | ml.c6gn.large | ml.c6gn.xlarge | ml.c6gn.2xlarge | ml.c6gn.4xlarge | ml.c6gn.8xlarge | ml.c6gn.12xlarge | ml.c6gn.16xlarge | ml.r6g.large | ml.r6g.xlarge | ml.r6g.2xlarge | ml.r6g.4xlarge | ml.r6g.8xlarge | ml.r6g.12xlarge | ml.r6g.16xlarge | ml.r6gd.large | ml.r6gd.xlarge | ml.r6gd.2xlarge | ml.r6gd.4xlarge | ml.r6gd.8xlarge | ml.r6gd.12xlarge | ml.r6gd.16xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.c8g.medium | ml.c8g.large | ml.c8g.xlarge | ml.c8g.2xlarge | ml.c8g.4xlarge | ml.c8g.8xlarge | ml.c8g.12xlarge | ml.c8g.16xlarge | ml.c8g.24xlarge | ml.c8g.48xlarge | ml.r7gd.medium | ml.r7gd.large | ml.r7gd.xlarge |

`ml.r7gd.2xlarge | ml.r7gd.4xlarge | ml.r7gd.8xlarge | ml.r7gd.12xlarge | ml.r7gd.16xlarge | ml.m8g.medium | ml.m8g.large | ml.m8g.xlarge | ml.m8g.2xlarge | ml.m8g.4xlarge | ml.m8g.8xlarge | ml.m8g.12xlarge | ml.m8g.16xlarge | ml.m8g.24xlarge | ml.m8g.48xlarge | ml.c6in.large | ml.c6in.xlarge | ml.c6in.2xlarge | ml.c6in.4xlarge | ml.c6in.8xlarge | ml.c6in.12xlarge | ml.c6in.16xlarge | ml.c6in.24xlarge | ml.c6in.32xlarge | ml.p6-b200.48xlarge | ml.p6-b300.48xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge`

Required: Yes

RecommendationId

The recommendation ID which uniquely identifies each recommendation.

Type: String

Required: Yes

Environment

The recommended environment variables to set in the model container for Real-Time Inference.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 1024.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 1024.

Value Pattern: `[\S\s]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

RecommendationJobCompiledOutputConfig

Service: Amazon SageMaker Service

Provides information about the output configuration for the compiled model.

Contents

S3OutputUri

Identifies the Amazon S3 bucket where you want SageMaker to store the compiled model artifacts.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RecommendationJobContainerConfig

Service: Amazon SageMaker Service

Specifies mandatory fields for running an Inference Recommender job directly in the CreateInferenceRecommendationsJob API. The fields specified in `ContainerConfig` override the corresponding fields in the model package. Use `ContainerConfig` if you want to specify these fields for the recommendation job but don't want to edit them in your model package.

Contents

DataInputConfig

Specifies the name and shape of the expected data inputs for your trained model with a JSON dictionary form. This field is used for optimizing your model using SageMaker Neo. For more information, see DataInputConfig.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[\S\s]+`

Required: No

Domain

The machine learning domain of the model and its components.

Valid Values: `COMPUTER_VISION | NATURAL_LANGUAGE_PROCESSING | MACHINE_LEARNING`

Type: String

Required: No

Framework

The machine learning framework of the container image.

Valid Values: `TENSORFLOW | PYTORCH | XGBOOST | SAGEMAKER-SCIKIT-LEARN`

Type: String

Required: No

FrameworkVersion

The framework version of the container image.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 10.

Pattern: `[0-9]\.[A-Za-z0-9.-]+`

Required: No

NearestModelName

The name of a pre-trained machine learning model benchmarked by Amazon SageMaker Inference Recommender that matches your model.

Valid Values: `efficientnetb7 | unet | xgboost | faster-rcnn-resnet101 | nasnetlarge | vgg16 | inception-v3 | mask-rcnn | sagemaker-scikit-learn | densenet201-gluon | resnet18v2-gluon | xception | densenet201 | yolov4 | resnet152 | bert-base-cased | xceptionV1-keras | resnet50 | retinanet`

Type: String

Required: No

PayloadConfig

Specifies the `SamplePayloadUrl` and all other sample payload-related fields.

Type: RecommendationJobPayloadConfig object

Required: No

SupportedEndpointType

The endpoint type to receive recommendations for. By default this is null, and the results of the inference recommendation job return a combined list of both real-time and serverless benchmarks. By specifying a value for this field, you can receive a longer list of benchmarks for the desired endpoint type.

Type: String

Valid Values: `RealTime | Serverless`

Required: No

SupportedInstanceTypes

A list of the instance types that are used to generate inferences in real-time.

Type: Array of strings

Required: No

SupportedResponseMIMETypes

The supported MIME types for the output data.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[-\w]+\/.+`

Required: No

Task

The machine learning task that the model accomplishes.

Valid Values: `IMAGE_CLASSIFICATION | OBJECT_DETECTION | TEXT_GENERATION | IMAGE_SEGMENTATION | FILL_MASK | CLASSIFICATION | REGRESSION | OTHER`

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RecommendationJobInferenceBenchmark

Service: Amazon SageMaker Service

The details for a specific benchmark from an Inference Recommender job.

Contents

ModelConfiguration

Defines the model configuration. Includes the specification name and environment parameters.

Type: ModelConfiguration object

Required: Yes

EndpointConfiguration

The endpoint configuration made by Inference Recommender during a recommendation job.

Type: EndpointOutputConfiguration object

Required: No

EndpointMetrics

The metrics for an existing endpoint compared in an Inference Recommender job.

Type: InferenceMetrics object

Required: No

FailureReason

The reason why a benchmark failed.

Type: String

Required: No

InvocationEndTime

A timestamp that shows when the benchmark completed.

Type: Timestamp

Required: No

InvocationStartTime

A timestamp that shows when the benchmark started.

Type: Timestamp

Required: No

Metrics

The metrics of recommendations.

Type: RecommendationMetrics object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RecommendationJobInputConfig

Service: Amazon SageMaker Service

The input configuration of the recommendation job.

Contents

ContainerConfig

Specifies mandatory fields for running an Inference Recommender job. The fields specified in `ContainerConfig` override the corresponding fields in the model package.

Type: RecommendationJobContainerConfig object

Required: No

EndpointConfigurations

Specifies the endpoint configuration to use for a job.

Type: Array of EndpointInputConfiguration objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

Endpoints

Existing customer endpoints on which to run an Inference Recommender job.

Type: Array of EndpointInfo objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

JobDurationInSeconds

Specifies the maximum duration of the job, in seconds. The maximum value is 18,000 seconds.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

ModelName

The name of the created model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

ModelPackageVersionArn

The Amazon Resource Name (ARN) of a versioned model package.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: No

ResourceLimit

Defines the resource limit of the job.

Type: RecommendationJobResourceLimit object

Required: No

TrafficPattern

Specifies the traffic pattern of the job.

Type: TrafficPattern object

Required: No

VolumeKmsKeyId

The Amazon Resource Name (ARN) of a AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt data on the storage volume attached to the ML compute

instance that hosts the endpoint. This key will be passed to SageMaker Hosting for endpoint creation.

The SageMaker execution role must have `kms:CreateGrant` permission in order to encrypt data on the storage volume of the endpoints created for inference recommendation. The inference recommendation job will fail asynchronously during endpoint configuration creation if the role passed does not have `kms:CreateGrant` permission.

The `KmsKeyId` can be any of the following formats:

- // KMS Key ID

  `"1234abcd-12ab-34cd-56ef-1234567890ab"`
- // Amazon Resource Name (ARN) of a KMS Key

  `"arn:aws:kms:<region>:<account>:key/<key-id-12ab-34cd-56ef-1234567890ab>"`
- // KMS Key Alias

  `"alias/ExampleAlias"`
- // Amazon Resource Name (ARN) of a KMS Key Alias

  `"arn:aws:kms:<region>:<account>:alias/<ExampleAlias>"`

For more information about key identifiers, see Key identifiers (KeyID) in the AWS Key Management Service (AWS KMS) documentation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

VpcConfig

Inference Recommender provisions SageMaker endpoints with access to VPC in the inference recommendation job.

Type: RecommendationJobVpcConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RecommendationJobOutputConfig

Service: Amazon SageMaker Service

Provides information about the output configuration for the compiled model.

Contents

CompiledOutputConfig

Provides information about the output configuration for the compiled model.

Type: RecommendationJobCompiledOutputConfig object

Required: No

KmsKeyId

The Amazon Resource Name (ARN) of a AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt your output artifacts with Amazon S3 server-side encryption. The SageMaker execution role must have `kms:GenerateDataKey` permission.

The `KmsKeyId` can be any of the following formats:

- // KMS Key ID

  `"1234abcd-12ab-34cd-56ef-1234567890ab"`

- // Amazon Resource Name (ARN) of a KMS Key

  `"arn:aws:kms:<region>:<account>:key/<key-id-12ab-34cd-56ef-1234567890ab>"`

- // KMS Key Alias

  `"alias/ExampleAlias"`

- // Amazon Resource Name (ARN) of a KMS Key Alias

  `"arn:aws:kms:<region>:<account>:alias/<ExampleAlias>"`

For more information about key identifiers, see Key identifiers (KeyID) in the AWS Key Management Service (AWS KMS) documentation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RecommendationJobPayloadConfig

Service: Amazon SageMaker Service

The configuration for the payload for a recommendation job.

Contents

SamplePayloadUrl

The Amazon Simple Storage Service (Amazon S3) path where the sample payload is stored. This path must point to a single gzip compressed tar archive (.tar.gz suffix).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

SupportedContentTypes

The supported MIME types for the input data.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RecommendationJobResourceLimit

Service: Amazon SageMaker Service

Specifies the maximum number of jobs that can run in parallel and the maximum number of jobs that can run.

Contents

MaxNumberOfTests

Defines the maximum number of load tests.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

MaxParallelOfTests

Defines the maximum number of parallel load tests.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RecommendationJobStoppingConditions

Service: Amazon SageMaker Service

Specifies conditions for stopping a job. When a job reaches a stopping condition limit, SageMaker ends the job.

Contents

FlatInvocations

Stops a load test when the number of invocations (TPS) peaks and flattens, which means that the instance has reached capacity. The default value is `Stop`. If you want the load test to continue after invocations have flattened, set the value to `Continue`.

Type: String

Valid Values: `Continue | Stop`

Required: No

MaxInvocations

The maximum number of requests per minute expected for the endpoint.

Type: Integer

Required: No

ModelLatencyThresholds

The interval of time taken by a model to respond as viewed from SageMaker. The interval includes the local communication time taken to send the request and to fetch the response from the container of a model and the time taken to complete the inference in the container.

Type: Array of ModelLatencyThreshold objects

Array Members: Fixed number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RecommendationJobVpcConfig

Service: Amazon SageMaker Service

Inference Recommender provisions SageMaker endpoints with access to VPC in the inference recommendation job.

Contents

SecurityGroupIds

The VPC security group IDs. IDs have the form of `sg-xxxxxxxx`. Specify the security groups for the VPC that is specified in the `Subnets` field.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: Yes

Subnets

The ID of the subnets in the VPC to which you want to connect your model.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

RecommendationMetrics

Service: Amazon SageMaker Service

The metrics of recommendations.

Contents

CostPerHour

Defines the cost per hour for the instance.

Type: Float

Required: No

CostPerInference

Defines the cost per inference for the instance .

Type: Float

Required: No

CpuUtilization

The expected CPU utilization at maximum invocations per minute for the instance.

NaN indicates that the value is not available.

Type: Float

Valid Range: Minimum value of 0.0.

Required: No

MaxInvocations

The expected maximum number of requests per minute for the instance.

Type: Integer

Required: No

MemoryUtilization

The expected memory utilization at maximum invocations per minute for the instance.

NaN indicates that the value is not available.

Type: Float

Valid Range: Minimum value of 0.0.

Required: No

ModelLatency

The expected model latency at maximum invocation per minute for the instance.

Type: Integer

Required: No

ModelSetupTime

The time it takes to launch new compute resources for a serverless endpoint. The time can vary depending on the model size, how long it takes to download the model, and the start-up time of the container.

NaN indicates that the value is not available.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RedshiftDatasetDefinition

Service: Amazon SageMaker Service

Configuration for Redshift Dataset Definition input.

Contents

ClusterId

The Redshift cluster Identifier.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `.*`

Required: Yes

ClusterRoleArn

The IAM role attached to your Redshift cluster that Amazon SageMaker uses to generate datasets.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

Database

The name of the Redshift database used in Redshift query execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.*`

Required: Yes

DbUser

The database user name used in Redshift query execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `.*`

Required: Yes

OutputFormat

The data storage format for Redshift query results.

Type: String

Valid Values: `PARQUET | CSV`

Required: Yes

OutputS3Uri

The location in Amazon S3 where the Redshift query results are stored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

QueryString

The SQL query statements to be executed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 4096.

Pattern: `[\s\S]+`

Required: Yes

KmsKeyId

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt data from a Redshift execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

OutputCompression

The compression used for Redshift query results.

Type: String

Valid Values: `None | GZIP | BZIP2 | ZSTD | SNAPPY`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RegisterModelStepMetadata

Service: Amazon SageMaker Service

Metadata for a register model job step.

Contents

Arn

The Amazon Resource Name (ARN) of the model package.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RemoteDebugConfig

Service: Amazon SageMaker Service

Configuration for remote debugging for the CreateTrainingJob API. To learn more about the remote debugging functionality of SageMaker, see Access a training container through AWS Systems Manager (SSM) for remote debugging.

Contents

EnableRemoteDebug

If set to True, enables remote debugging.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RemoteDebugConfigForUpdate

Service: Amazon SageMaker Service

Configuration for remote debugging for the UpdateTrainingJob API. To learn more about the remote debugging functionality of SageMaker, see Access a training container through AWS Systems Manager (SSM) for remote debugging.

Contents

EnableRemoteDebug

If set to True, enables remote debugging.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RenderableTask

Service: Amazon SageMaker Service

Contains input values for a task.

Contents

Input

A JSON object that contains values for the variables defined in the template. It is made available to the template under the substitution variable `task.input`. For example, if you define a variable `task.input.text` in your template, you can supply the variable in the JSON object as `"text": "sample text"`.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 128000.

Pattern: `[\S\s]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RenderingError

Service: Amazon SageMaker Service

A description of an error that occurred while rendering the template.

Contents

Code

A unique identifier for a specific class of errors.

Type: String

Required: Yes

Message

A human-readable message describing the error.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RepositoryAuthConfig

Service: Amazon SageMaker Service

Specifies an authentication configuration for the private docker registry where your model image is hosted. Specify a value for this property only if you specified `Vpc` as the value for the `RepositoryAccessMode` field of the `ImageConfig` object that you passed to a call to `CreateModel` and the private Docker registry where the model image is hosted requires authentication.

Contents

RepositoryCredentialsProviderArn

The Amazon Resource Name (ARN) of an AWS Lambda function that provides credentials to authenticate to the private Docker registry where your model image is hosted. For information about how to create an AWS Lambda function, see Create a Lambda function with the console in the *AWS Lambda Developer Guide*.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ReservedCapacityOffering

Service: Amazon SageMaker Service

Details about a reserved capacity offering for a training plan offering.

For more information about how to reserve GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Contents

InstanceCount

The number of instances in the reserved capacity offering.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 256.

Required: Yes

InstanceType

The instance type for the reserved capacity offering.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.trn1.32xlarge | ml.trn2.48xlarge | ml.p6-b200.48xlarge | ml.p4de.24xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge | ml.p6-b300.48xlarge`

Required: Yes

AvailabilityZone

The availability zone for the reserved capacity offering.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-z]+\-[0-9a-z\-]+`

Required: No

DurationHours

The number of whole hours in the total duration for this reserved capacity offering.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 87600.

Required: No

DurationMinutes

The additional minutes beyond whole hours in the total duration for this reserved capacity offering.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 59.

Required: No

EndTime

The end time of the reserved capacity offering.

Type: Timestamp

Required: No

ExtensionEndTime

The end time of the extension for the reserved capacity offering.

Type: Timestamp

Required: No

ExtensionStartTime

The start time of the extension for the reserved capacity offering.

Type: Timestamp

Required: No

ReservedCapacityType

The type of reserved capacity offering.

Type: String

Valid Values: `UltraServer | Instance`

Required: No

StartTime

The start time of the reserved capacity offering.

Type: Timestamp

Required: No

UltraServerCount

The number of UltraServers included in this reserved capacity offering.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

UltraServerType

The type of UltraServer included in this reserved capacity offering, such as ml.u-p6e-gb200x72.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `ml.[a-z0-9\-.]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ReservedCapacitySummary

Service: Amazon SageMaker Service

Details of a reserved capacity for the training plan.

For more information about how to reserve GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Contents

InstanceType

The instance type for the reserved capacity.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.trn1.32xlarge | ml.trn2.48xlarge | ml.p6-b200.48xlarge | ml.p4de.24xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge | ml.p6-b300.48xlarge`

Required: Yes

ReservedCapacityArn

The Amazon Resource Name (ARN); of the reserved capacity.

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:reserved-capacity/.*`

Required: Yes

Status

The current status of the reserved capacity.

Type: String

Valid Values: `Pending | Active | Scheduled | Expired | Failed`

Required: Yes

TotalInstanceCount

The total number of instances in the reserved capacity.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

AvailabilityZone

The availability zone for the reserved capacity.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-z]+\-[0-9a-z\-]+`

Required: No

AvailabilityZoneId

The Availability Zone ID of the reserved capacity.

Type: String

Pattern: `[a-z]{3}\d-az\d`

Required: No

DurationHours

The number of whole hours in the total duration for this reserved capacity.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 87600.

Required: No

DurationMinutes

The additional minutes beyond whole hours in the total duration for this reserved capacity.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 59.

Required: No

EndTime

The end time of the reserved capacity.

Type: Timestamp

Required: No

ReservedCapacityType

The type of reserved capacity.

Type: String

Valid Values: `UltraServer | Instance`

Required: No

StartTime

The start time of the reserved capacity.

Type: Timestamp

Required: No

UltraServerCount

The number of UltraServers included in this reserved capacity.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

UltraServerType

The type of UltraServer included in this reserved capacity, such as ml.u-p6e-gb200x72.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `ml.[a-z0-9\-.]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResolvedAttributes

Service: Amazon SageMaker Service

The resolved attributes.

Contents

AutoMLJobObjective

Specifies a metric to minimize or maximize as the objective of an AutoML job.

Type: AutoMLJobObjective object

Required: No

CompletionCriteria

How long a job is allowed to run, or how many candidates a job is allowed to generate.

Type: AutoMLJobCompletionCriteria object

Required: No

ProblemType

The problem type.

Type: String

Valid Values: `BinaryClassification | MulticlassClassification | Regression`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResourceCatalog

Service: Amazon SageMaker Service

A resource catalog containing all of the resources of a specific resource type within a resource owner account. For an example on sharing the Amazon SageMaker Feature Store `DefaultFeatureGroupCatalog`, see Share Amazon SageMaker Catalog resource type in the Amazon SageMaker Developer Guide.

Contents

CreationTime

The time the `ResourceCatalog` was created.

Type: Timestamp

Required: Yes

Description

A free form description of the `ResourceCatalog`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

ResourceCatalogArn

The Amazon Resource Name (ARN) of the `ResourceCatalog`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:sagemaker-catalog/.*`

Required: Yes

ResourceCatalogName

The name of the `ResourceCatalog`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResourceConfig

Service: Amazon SageMaker Service

Describes the resources, including machine learning (ML) compute instances and ML storage volumes, to use for model training.

Contents

InstanceCount

The number of ML compute instances to use. For distributed training, provide a value greater than 1.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

InstanceGroups

The configuration of a heterogeneous cluster in JSON format.

Type: Array of InstanceGroup objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

InstancePlacementConfig

Configuration for how training job instances are placed and allocated within UltraServers. Only applicable for UltraServer capacity.

Type: InstancePlacementConfig object

Required: No

InstanceType

The ML compute instance type.

Type: String

Valid Values: `ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge`

| ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p3dn.24xlarge | ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5n.xlarge | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.8xlarge | ml.c6i.4xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.8xlarge | ml.r5d.12xlarge | ml.r5d.16xlarge | ml.r5d.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.p6-b200.48xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge | ml.p6-b300.48xlarge | ml.g7e.2xlarge | ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge | ml.g7e.48xlarge

Required: No

KeepAlivePeriodInSeconds

The duration of time in seconds to retain configured resources in a warm pool for subsequent training jobs.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 21600.

Required: No

TrainingPlanArn

The Amazon Resource Name (ARN); of the training plan to use for this resource configuration.

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Required: No

VolumeKmsKeyId

The AWS KMS key that SageMaker uses to encrypt data on the storage volume attached to the ML compute instance(s) that run the training job.

> **Note**
>
> Certain Nitro-based instances include local storage, dependent on the instance type. Local storage volumes are encrypted using a hardware module on the instance. You can't request a `VolumeKmsKeyId` when using an instance type with local storage.
> For a list of instance types that support local instance storage, see Instance Store Volumes.
> For more information about local instance storage encryption, see SSD Instance Store Volumes.

The `VolumeKmsKeyId` can be in any of the following formats:

- // KMS Key ID

  `"1234abcd-12ab-34cd-56ef-1234567890ab"`

- // Amazon Resource Name (ARN) of a KMS Key

  ```
  "arn:aws:kms:us-
  west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab"
  ```

  Type: String

  Length Constraints: Minimum length of 0. Maximum length of 2048.

  Pattern: `[a-zA-Z0-9:/_-]*`

  Required: No

VolumeSizeInGB

The size of the ML storage volume that you want to provision.

SageMaker automatically selects the volume size for serverless training jobs. You cannot customize this setting.

ML storage volumes store model artifacts and incremental states. Training algorithms might also use the ML storage volume for scratch space. If you want to store the training data in the ML storage volume, choose `File` as the `TrainingInputMode` in the algorithm specification.

When using an ML instance with NVMe SSD volumes, SageMaker doesn't provision Amazon EBS General Purpose SSD (gp2) storage. Available storage is fixed to the NVMe-type instance's storage capacity. SageMaker configures storage paths for training datasets, checkpoints, model artifacts, and outputs to use the entire capacity of the instance storage. For example, ML instance families with the NVMe-type instance storage include `ml.p4d`, `ml.g4dn`, and `ml.g5`.

When using an ML instance with the EBS-only storage option and without instance storage, you must define the size of EBS volume through `VolumeSizeInGB` in the `ResourceConfig` API. For example, ML instance families that use EBS volumes include `ml.c5` and `ml.p2`.

To look up instance types and their instance storage types and volumes, see Amazon EC2 Instance Types.

To find the default local paths defined by the SageMaker training platform, see Amazon SageMaker Training Storage Folders for Training Datasets, Checkpoints, Model Artifacts, and Outputs.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResourceConfigForUpdate

Service: Amazon SageMaker Service

The `ResourceConfig` to update `KeepAlivePeriodInSeconds`. Other fields in the `ResourceConfig` cannot be updated.

Contents

KeepAlivePeriodInSeconds

The `KeepAlivePeriodInSeconds` value specified in the `ResourceConfig` to update.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 21600.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResourceLimits

Service: Amazon SageMaker Service

Specifies the maximum number of training jobs and parallel training jobs that a hyperparameter tuning job can launch.

Contents

MaxParallelTrainingJobs

The maximum number of concurrent training jobs that a hyperparameter tuning job can launch.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

MaxNumberOfTrainingJobs

The maximum number of training jobs that a hyperparameter tuning job can launch.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

MaxRuntimeInSeconds

The maximum time in seconds that a hyperparameter tuning job can run.

Type: Integer

Valid Range: Minimum value of 120. Maximum value of 15768000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResourceSharingConfig

Service: Amazon SageMaker Service

Resource sharing configuration.

Contents

Strategy

The strategy of how idle compute is shared within the cluster. The following are the options of strategies.

- `DontLend`: entities do not lend idle compute.
- `Lend`: entities can lend idle compute to entities that can borrow.
- `LendandBorrow`: entities can lend idle compute and borrow idle compute from other entities.

Default is `LendandBorrow`.

Type: String

Valid Values: `Lend | DontLend | LendAndBorrow`

Required: Yes

AbsoluteBorrowLimits

The absolute limits on compute resources that can be borrowed from idle compute. When specified, these limits define the maximum amount of specific resource types (such as accelerators, vCPU, or memory) that an entity can borrow, regardless of the percentage-based `BorrowLimit`.

Type: Array of ComputeQuotaResourceConfig objects

Array Members: Minimum number of 0 items. Maximum number of 15 items.

Required: No

BorrowLimit

The limit on how much idle compute can be borrowed.The values can be 1 - 500 percent of idle compute that the team is allowed to borrow.

Default is 50.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResourceSpec

Service: Amazon SageMaker Service

Specifies the ARN's of a SageMaker AI image and SageMaker AI image version, and the instance type that the version runs on.

> **Note**
>
> When both `SageMakerImageVersionArn` and `SageMakerImageArn` are passed, `SageMakerImageVersionArn` is used. Any updates to `SageMakerImageArn` will not take effect if `SageMakerImageVersionArn` already exists in the `ResourceSpec` because `SageMakerImageVersionArn` always takes precedence. To clear the value set for `SageMakerImageVersionArn`, pass None as the value.

Contents

InstanceType

The instance type that the image version runs on.

> **Note**
>
> **JupyterServer apps** only support the `system` value.
> For **KernelGateway apps**, the `system` value is translated to `ml.t3.medium`.
> KernelGateway apps also support all other values for available instance types.

Type: String

Valid Values: `system | ml.t3.micro | ml.t3.small | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.8xlarge | ml.m5d.12xlarge | ml.m5d.16xlarge | ml.m5d.24xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p3dn.24xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge |`

ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.geospatial.interactive | ml.p4d.24xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.p5.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.m6id.large | ml.m6id.xlarge | ml.m6id.2xlarge | ml.m6id.4xlarge | ml.m6id.8xlarge | ml.m6id.12xlarge | ml.m6id.16xlarge | ml.m6id.24xlarge | ml.m6id.32xlarge | ml.c6id.large | ml.c6id.xlarge | ml.c6id.2xlarge | ml.c6id.4xlarge | ml.c6id.8xlarge | ml.c6id.12xlarge | ml.c6id.16xlarge | ml.c6id.24xlarge | ml.c6id.32xlarge | ml.r6id.large | ml.r6id.xlarge | ml.r6id.2xlarge | ml.r6id.4xlarge | ml.r6id.8xlarge | ml.r6id.12xlarge | ml.r6id.16xlarge | ml.r6id.24xlarge | ml.r6id.32xlarge | ml.p5.4xlarge

Required: No

LifecycleConfigArn

The Amazon Resource Name (ARN) of the Lifecycle Configuration attached to the Resource.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:studio-lifecycle-config/.*|None)`

Required: No

SageMakerImageArn

The ARN of the SageMaker AI image that the image version belongs to.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image/[a-zA-Z0-9]([-.]?[a-zA-Z0-9])*`

Required: No

SageMakerImageVersionAlias

The SageMakerImageVersionAlias of the image to launch with. This value is in SemVer 2.0.0 versioning format.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(^\d+$)|(^\d+.\d+$)|(^(0|[1-9]\d*)\.(0|[1-9]\d*)\.(0|[1-9]\d*)(?:-((?:0|[1-9]\d*|\d*[a-zA-Z-][0-9a-zA-Z-]*)(?:\.(?:0|[1-9]\d*|\d*[a-zA-Z-][0-9a-zA-Z-]*))*))?(?:\+([0-9a-zA-Z-]+(?:\.[0-9a-zA-Z-]+)*))?$)`

Required: No

SageMakerImageVersionArn

The ARN of the image version created on the instance. To clear the value set for `SageMakerImageVersionArn`, pass `None` as the value.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws(-[\w]+)*:sagemaker:.+:[0-9]{12}:image-version/[a-z0-9]([-.]?[a-z0-9])*/[0-9]+|None)`

Required: No

TrainingPlanArn

The ARN of the SageMaker AI Training Plan to use for this app. When you specify a training plan, the app launches on reserved GPU capacity. This field is supported for JupyterLab and CodeEditor app types.

For more information about how to reserve GPU capacity with SageMaker AI Training Plans, see Using training plans in Studio applications.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*|None)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetentionPolicy

Service: Amazon SageMaker Service

The retention policy for data stored on an Amazon Elastic File System volume.

Contents

HomeEfsFileSystem

The default is `Retain`, which specifies to keep the data stored on the Amazon EFS volume.

Specify `Delete` to delete the data stored on the Amazon EFS volume.

Type: String

Valid Values: `Retain | Delete`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetryStrategy

Service: Amazon SageMaker Service

The retry strategy to use when a training job fails due to an `InternalServerError`. `RetryStrategy` is specified as part of the `CreateTrainingJob` and `CreateHyperParameterTuningJob` requests. You can add the `StoppingCondition` parameter to the request to limit the training time for the complete job.

Contents

MaximumRetryAttempts

The number of times to retry the job. When the job is retried, it's `SecondaryStatus` is changed to `STARTING`.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 30.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RoleGroupAssignment

Service: Amazon SageMaker Service

Defines the mapping between an in-app role and the AWS IAM Identity Center group patterns that should be assigned to that role within the SageMaker Partner AI App.

Contents

GroupPatterns

A list of AWS IAM Identity Center group patterns that should be assigned to the specified role. Group patterns support wildcard matching using *.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\w+=,.@*-]+`

Required: Yes

RoleName

The name of the in-app role within the SageMaker Partner AI App. The specific roles available depend on the app type and version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(?!\s*$).+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

RollingDeploymentPolicy

Service: Amazon SageMaker Service

The configurations that SageMaker uses when updating the AMI versions.

Contents

MaximumBatchSize

The maximum amount of instances in the cluster that SageMaker can update at a time.

Type: CapacitySizeConfig object

Required: Yes

RollbackMaximumBatchSize

The maximum amount of instances in the cluster that SageMaker can roll back at a time.

Type: CapacitySizeConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RollingUpdatePolicy

Service: Amazon SageMaker Service

Specifies a rolling deployment strategy for updating a SageMaker endpoint.

Contents

MaximumBatchSize

Batch size for each rolling step to provision capacity and turn on traffic on the new endpoint fleet, and terminate capacity on the old endpoint fleet. Value must be between 5% to 50% of the variant's total instance count.

Type: CapacitySize object

Required: Yes

WaitIntervalInSeconds

The length of the baking period, during which SageMaker monitors alarms for each batch on the new fleet.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 3600.

Required: Yes

MaximumExecutionTimeoutInSeconds

The time limit for the total deployment. Exceeding this limit causes a timeout.

Type: Integer

Valid Range: Minimum value of 600. Maximum value of 28800.

Required: No

RollbackMaximumBatchSize

Batch size for rollback to the old endpoint fleet. Each rolling step to provision capacity and turn on traffic on the old endpoint fleet, and terminate capacity on the new endpoint fleet. If this field is absent, the default value will be set to 100% of total capacity which means to bring up the whole capacity of the old fleet at once during rollback.

Type: CapacitySize object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RSessionAppSettings

Service: Amazon SageMaker Service

A collection of settings that apply to an `RSessionGateway` app.

Contents

CustomImages

A list of custom SageMaker AI images that are configured to run as a RSession app.

Type: Array of CustomImage objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

DefaultResourceSpec

Specifies the ARN's of a SageMaker AI image and SageMaker AI image version, and the instance type that the version runs on.

> **Note**
>
> When both `SageMakerImageVersionArn` and `SageMakerImageArn` are passed, `SageMakerImageVersionArn` is used. Any updates to `SageMakerImageArn` will not take effect if `SageMakerImageVersionArn` already exists in the `ResourceSpec` because `SageMakerImageVersionArn` always takes precedence. To clear the value set for `SageMakerImageVersionArn`, pass `None` as the value.

Type: ResourceSpec object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

RStudioServerProAppSettings

Service: Amazon SageMaker Service

A collection of settings that configure user interaction with the `RStudioServerPro` app.

Contents

AccessStatus

Indicates whether the current user has access to the `RStudioServerPro` app.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

UserGroup

The level of permissions that the user has within the `RStudioServerPro` app. This value defaults to `User`. The `Admin` value allows the user access to the RStudio Administrative Dashboard.

Type: String

Valid Values: `R_STUDIO_ADMIN | R_STUDIO_USER`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RStudioServerProDomainSettings

Service: Amazon SageMaker Service

A collection of settings that configure the `RStudioServerPro` Domain-level app.

Contents

DomainExecutionRoleArn

The ARN of the execution role for the `RStudioServerPro` Domain-level app.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

DefaultResourceSpec

Specifies the ARN's of a SageMaker AI image and SageMaker AI image version, and the instance type that the version runs on.

> **Note**
>
> When both `SageMakerImageVersionArn` and `SageMakerImageArn` are passed, `SageMakerImageVersionArn` is used. Any updates to `SageMakerImageArn` will not take effect if `SageMakerImageVersionArn` already exists in the `ResourceSpec` because `SageMakerImageVersionArn` always takes precedence. To clear the value set for `SageMakerImageVersionArn`, pass `None` as the value.

Type: ResourceSpec object

Required: No

RStudioConnectUrl

A URL pointing to an RStudio Connect server.

Type: String

Required: No

RStudioPackageManagerUrl

A URL pointing to an RStudio Package Manager server.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RStudioServerProDomainSettingsForUpdate

Service: Amazon SageMaker Service

A collection of settings that update the current configuration for the `RStudioServerPro` Domain-level app.

Contents

DomainExecutionRoleArn

The execution role for the `RStudioServerPro` Domain-level app.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: Yes

DefaultResourceSpec

Specifies the ARN's of a SageMaker AI image and SageMaker AI image version, and the instance type that the version runs on.

> **Note**
>
> When both `SageMakerImageVersionArn` and `SageMakerImageArn` are passed, `SageMakerImageVersionArn` is used. Any updates to `SageMakerImageArn` will not take effect if `SageMakerImageVersionArn` already exists in the `ResourceSpec` because `SageMakerImageVersionArn` always takes precedence. To clear the value set for `SageMakerImageVersionArn`, pass `None` as the value.

Type: ResourceSpec object

Required: No

RStudioConnectUrl

A URL pointing to an RStudio Connect server.

Type: String

Required: No

RStudioPackageManagerUrl

A URL pointing to an RStudio Package Manager server.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3DataSource

Service: Amazon SageMaker Service

Describes the S3 data source.

Your input bucket must be in the same AWS region as your training job.

Contents

S3DataType

If you choose `S3Prefix`, `S3Uri` identifies a key name prefix. SageMaker uses all objects that match the specified key name prefix for model training.

If you choose `ManifestFile`, `S3Uri` identifies an object that is a manifest file containing a list of object keys that you want SageMaker to use for model training.

If you choose `AugmentedManifestFile`, `S3Uri` identifies an object that is an augmented manifest file in JSON lines format. This file contains the data you want to use for model training. `AugmentedManifestFile` can only be used if the Channel's input mode is `Pipe`.

If you choose `Converse`, `S3Uri` identifies an Amazon S3 location that contains data formatted according to Converse format. This format structures conversational messages with specific roles and content types used for training and fine-tuning foundational models.

Type: String

Valid Values: `ManifestFile | S3Prefix | AugmentedManifestFile | Converse`

Required: Yes

S3Uri

Depending on the value specified for the `S3DataType`, identifies either a key name prefix or a manifest. For example:

- A key name prefix might look like this: `s3://bucketname/exampleprefix/`
- A manifest might look like this: `s3://bucketname/example.manifest`

  A manifest is an S3 object which is a JSON file consisting of an array of elements. The first element is a prefix which is followed by one or more suffixes. SageMaker appends the suffix elements to the prefix to get a full set of `S3Uri`. Note that the prefix must be a valid non-

empty `S3Uri` that precludes users from specifying a manifest whose individual `S3Uri` is sourced from different S3 buckets.

The following code example shows a valid manifest format:

```
[ {"prefix": "s3://customer_bucket/some/prefix/"},

"relative/path/to/custdata-1",

"relative/path/custdata-2",

...

"relative/path/custdata-N"

]
```

This JSON is equivalent to the following `S3Uri` list:

```
s3://customer_bucket/some/prefix/relative/path/to/custdata-1

s3://customer_bucket/some/prefix/relative/path/custdata-2

...

s3://customer_bucket/some/prefix/relative/path/custdata-N
```

The complete set of `S3Uri` in this manifest is the input data for the channel for this data source. The object that each `S3Uri` points to must be readable by the IAM role that SageMaker uses to perform tasks on your behalf.

Your input bucket must be located in same AWS region as your training job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

AttributeNames

A list of one or more attribute names to use that are found in a specified augmented manifest file.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 16 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.+`

Required: No

HubAccessConfig

The configuration for a private hub model reference that points to a SageMaker JumpStart public hub model.

Type: HubAccessConfig object

Required: No

InstanceGroupNames

A list of names of instance groups that get data from the S3 data source.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `.+`

Required: No

ModelAccessConfig

The access configuration file to control access to the ML model. You can explicitly accept the model end-user license agreement (EULA) within the `ModelAccessConfig`.

- If you are a Jumpstart user, see the End-user license agreements section for more details on accepting the EULA.

- If you are an AutoML user, see the *Optional Parameters* section of *Create an AutoML job to fine-tune text generation models using the API* for details on How to set the EULA acceptance when fine-tuning a model using the AutoML API.

Type: ModelAccessConfig object

Required: No

S3DataDistributionType

If you want SageMaker to replicate the entire dataset on each ML compute instance that is launched for model training, specify `FullyReplicated`.

If you want SageMaker to replicate a subset of data on each ML compute instance that is launched for model training, specify `ShardedByS3Key`. If there are *n* ML compute instances launched for a training job, each instance gets approximately 1/*n* of the number of S3 objects. In this case, model training on each machine uses only the subset of training data.

Don't choose more ML compute instances for training than available S3 objects. If you do, some nodes won't get any data and you will pay for nodes that aren't getting any training data. This applies in both File and Pipe modes. Keep this in mind when developing algorithms.

In distributed training, where you use multiple ML compute EC2 instances, you might choose `ShardedByS3Key`. If the algorithm requires copying training data to the ML storage volume (when `TrainingInputMode` is set to `File`), this copies 1/*n* of the number of objects.

Type: String

Valid Values: `FullyReplicated | ShardedByS3Key`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3FileSystem

Service: Amazon SageMaker Service

A custom file system in Amazon S3. This is only supported in Amazon SageMaker Unified Studio.

Contents

S3Uri

The Amazon S3 URI that specifies the location in S3 where files are stored, which is mounted within the Studio environment. For example: `s3://<bucket-name>/<prefix>/`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(s3)://([^/]+)/?(.*)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3FileSystemConfig

Service: Amazon SageMaker Service

Configuration for the custom Amazon S3 file system.

Contents

S3Uri

The Amazon S3 URI of the S3 file system configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(s3)://([^/]+)/?(.*)`

Required: Yes

MountPath

The file system path where the Amazon S3 storage location will be mounted within the Amazon SageMaker Studio environment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3ModelDataSource

Service: Amazon SageMaker Service

Specifies the S3 location of ML model data to deploy.

Contents

CompressionType

Specifies how the ML model data is prepared.

If you choose `Gzip` and choose `S3Object` as the value of `S3DataType`, `S3Uri` identifies an object that is a gzip-compressed TAR archive. SageMaker will attempt to decompress and untar the object during model deployment.

If you choose `None` and chooose `S3Object` as the value of `S3DataType`, `S3Uri` identifies an object that represents an uncompressed ML model to deploy.

If you choose None and choose `S3Prefix` as the value of `S3DataType`, `S3Uri` identifies a key name prefix, under which all objects represents the uncompressed ML model to deploy.

If you choose None, then SageMaker will follow rules below when creating model data files under /opt/ml/model directory for use by your inference code:

- If you choose `S3Object` as the value of `S3DataType`, then SageMaker will split the key of the S3 object referenced by `S3Uri` by slash (/), and use the last part as the filename of the file holding the content of the S3 object.
- If you choose `S3Prefix` as the value of `S3DataType`, then for each S3 object under the key name pefix referenced by `S3Uri`, SageMaker will trim its key by the prefix, and use the remainder as the path (relative to `/opt/ml/model`) of the file holding the content of the S3 object. SageMaker will split the remainder by slash (/), using intermediate parts as directory names and the last part as filename of the file holding the content of the S3 object.
- Do not use any of the following as file names or directory names:
  - An empty or blank string
  - A string which contains null bytes
  - A string longer than 255 bytes
  - A single dot (`.`)
  - A double dot (`..`)

- Ambiguous file names will result in model deployment failure. For example, if your uncompressed ML model consists of two S3 objects `s3://mybucket/model/weights` and `s3://mybucket/model/weights/part1` and you specify `s3://mybucket/model/` as the value of `S3Uri` and `S3Prefix` as the value of `S3DataType`, then it will result in name clash between `/opt/ml/model/weights` (a regular file) and `/opt/ml/model/weights/` (a directory).
- Do not organize the model artifacts in S3 console using folders. When you create a folder in S3 console, S3 creates a 0-byte object with a key set to the folder name you provide. They key of the 0-byte object ends with a slash (/) which violates SageMaker restrictions on model artifact file names, leading to model deployment failure.

Type: String

Valid Values: `None | Gzip`

Required: Yes

S3DataType

Specifies the type of ML model data to deploy.

If you choose `S3Prefix`, `S3Uri` identifies a key name prefix. SageMaker uses all objects that match the specified key name prefix as part of the ML model data to deploy. A valid key name prefix identified by `S3Uri` always ends with a forward slash (/).

If you choose `S3Object`, `S3Uri` identifies an object that is the ML model data to deploy.

Type: String

Valid Values: `S3Prefix | S3Object`

Required: Yes

S3Uri

Specifies the S3 path of ML model data to deploy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

ETag

The ETag associated with S3 URI.

Type: String

Required: No

HubAccessConfig

Configuration information for hub access.

Type: InferenceHubAccessConfig object

Required: No

ManifestEtag

The ETag associated with Manifest S3 URI.

Type: String

Required: No

ManifestS3Uri

The Amazon S3 URI of the manifest file. The manifest file is a CSV file that stores the artifact locations.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

ModelAccessConfig

Specifies the access configuration file for the ML model. You can explicitly accept the model end-user license agreement (EULA) within the `ModelAccessConfig`. You are responsible for reviewing and complying with any applicable license terms and making sure they are acceptable for your use case before downloading or using a model.

Type: ModelAccessConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3Presign

Service: Amazon SageMaker Service

This object defines the access restrictions to Amazon S3 resources that are included in custom worker task templates using the Liquid filter, `grant_read_access`.

To learn more about how custom templates are created, see Create custom worker task templates.

Contents

IamPolicyConstraints

Use this parameter to specify the allowed request source. Possible sources are either `SourceIp` or `VpcSourceIp`.

Type: IamPolicyConstraints object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3StorageConfig

Service: Amazon SageMaker Service

The Amazon Simple Storage (Amazon S3) location and security configuration for `OfflineStore`.

Contents

S3Uri

The S3 URI, or location in Amazon S3, of `OfflineStore`.

S3 URIs have a format similar to the following: `s3://example-bucket/prefix/`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

KmsKeyId

The AWS Key Management Service (KMS) key ARN of the key used to encrypt any objects written into the `OfflineStore` S3 location.

The IAM `roleARN` that is passed as a parameter to `CreateFeatureGroup` must have below permissions to the `KmsKeyId`:

- `"kms:GenerateDataKey"`

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

ResolvedOutputS3Uri

The S3 path where offline records are written.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ScalingPolicy

Service: Amazon SageMaker Service

An object containing a recommended scaling policy.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

TargetTracking

A target tracking scaling policy. Includes support for predefined or customized metrics.

Type: TargetTrackingScalingPolicyConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ScalingPolicyMetric

Service: Amazon SageMaker Service

The metric for a scaling policy.

Contents

InvocationsPerInstance

The number of invocations sent to a model, normalized by `InstanceCount` in each ProductionVariant. `1/numberOfInstances` is sent as the value on each request, where `numberOfInstances` is the number of active instances for the ProductionVariant behind the endpoint at the time of the request.

Type: Integer

Required: No

ModelLatency

The interval of time taken by a model to respond as viewed from SageMaker. This interval includes the local communication times taken to send the request and to fetch the response from the container of a model and the time taken to complete the inference in the container.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ScalingPolicyObjective

Service: Amazon SageMaker Service

An object where you specify the anticipated traffic pattern for an endpoint.

Contents

MaxInvocationsPerMinute

The maximum number of expected requests to your endpoint per minute.

Type: Integer

Required: No

MinInvocationsPerMinute

The minimum number of expected requests to your endpoint per minute.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ScheduleConfig

Service: Amazon SageMaker Service

Configuration details about the monitoring schedule.

Contents

ScheduleExpression

A cron expression that describes details about the monitoring schedule.

The supported cron expressions are:

- If you want to set the job to start every hour, use the following:

  `Hourly: cron(0 * ? * * *)`

- If you want to start the job daily:

  `cron(0 [00-23] ? * * *)`

- If you want to run the job one time, immediately, use the following keyword:

  `NOW`

For example, the following are valid cron expressions:

- Daily at noon UTC: `cron(0 12 ? * * *)`
- Daily at midnight UTC: `cron(0 0 ? * * *)`

To support running every 6, 12 hours, the following are also supported:

`cron(0 [00-23]/[01-24] ? * * *)`

For example, the following are valid cron expressions:

- Every 12 hours, starting at 5pm UTC: `cron(0 17/12 ? * * *)`
- Every two hours starting at midnight: `cron(0 0/2 ? * * *)`

> **Note**
>
> - Even though the cron expression is set to start at 5PM UTC, note that there could be a delay of 0-20 minutes from the actual requested time to run the execution.

- We recommend that if you would like a daily schedule, you do not provide this parameter. Amazon SageMaker AI will pick a time for running every day.

You can also specify the keyword `NOW` to run the monitoring job immediately, one time, without recurring.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: Yes

DataAnalysisEndTime

Sets the end time for a monitoring job window. Express this time as an offset to the times that you schedule your monitoring jobs to run. You schedule monitoring jobs with the `ScheduleExpression` parameter. Specify this offset in ISO 8601 duration format. For example, if you want to end the window one hour before the start of each monitoring job, you would specify: `"-PT1H"`.

The end time that you specify must not follow the start time that you specify by more than 24 hours. You specify the start time with the `DataAnalysisStartTime` parameter.

If you set `ScheduleExpression` to `NOW`, this parameter is required.

Type: String

Required: No

DataAnalysisStartTime

Sets the start time for a monitoring job window. Express this time as an offset to the times that you schedule your monitoring jobs to run. You schedule monitoring jobs with the `ScheduleExpression` parameter. Specify this offset in ISO 8601 duration format. For example, if you want to monitor the five hours of data in your dataset that precede the start of each monitoring job, you would specify: `"-PT5H"`.

The start time that you specify must not precede the end time that you specify by more than 24 hours. You specify the end time with the `DataAnalysisEndTime` parameter.

If you set `ScheduleExpression` to `NOW`, this parameter is required.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ScheduledUpdateConfig

Service: Amazon SageMaker Service

The configuration object of the schedule that SageMaker follows when updating the AMI.

Contents

ScheduleExpression

A cron expression that specifies the schedule that SageMaker follows when updating the AMI.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: Yes

DeploymentConfig

The configuration to use when updating the AMI versions.

Type: DeploymentConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SchedulerConfig

Service: Amazon SageMaker Service

Cluster policy configuration. This policy is used for task prioritization and fair-share allocation. This helps prioritize critical workloads and distributes idle compute across entities.

Contents

FairShare

When enabled, entities borrow idle compute based on their assigned `FairShareWeight`.

When disabled, entities borrow idle compute based on a first-come first-serve basis.

Default is `Enabled`.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

IdleResourceSharing

Configuration for sharing idle compute resources across entities in the cluster. When enabled, unallocated resources are automatically calculated and made available for entities to borrow.

Type: String

Valid Values: `Enabled | Disabled`

Required: No

PriorityClasses

List of the priority classes, `PriorityClass`, of the cluster policy. When specified, these class configurations define how tasks are queued.

Type: Array of PriorityClass objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SearchExpression

Service: Amazon SageMaker Service

A multi-expression that searches for the specified resource or resources in a search. All resource objects that satisfy the expression's condition are included in the search results. You must specify at least one subexpression, filter, or nested filter. A `SearchExpression` can contain up to twenty elements.

A `SearchExpression` contains the following components:

- A list of `Filter` objects. Each filter defines a simple Boolean expression comprised of a resource property name, Boolean operator, and value.
- A list of `NestedFilter` objects. Each nested filter defines a list of Boolean expressions using a list of resource properties. A nested filter is satisfied if a single object in the list satisfies all Boolean expressions.
- A list of `SearchExpression` objects. A search expression object can be nested in a list of search expression objects.
- A Boolean operator: `And` or `Or`.

Contents

Filters

A list of filter objects.

Type: Array of Filter objects

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Required: No

NestedFilters

A list of nested filter objects.

Type: Array of NestedFilters objects

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Required: No

Operator

A Boolean operator used to evaluate the search expression. If you want every conditional statement in all lists to be satisfied for the entire search expression to be true, specify And. If only a single conditional statement needs to be true for the entire search expression to be true, specify `Or`. The default value is And.

Type: String

Valid Values: `And | Or`

Required: No

SubExpressions

A list of search expression objects.

Type: Array of SearchExpression objects

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SearchRecord

Service: Amazon SageMaker Service

A single resource returned as part of the Search API response.

Contents

Endpoint

A hosted endpoint for real-time inference.

Type: Endpoint object

Required: No

Experiment

The properties of an experiment.

Type: Experiment object

Required: No

FeatureGroup

Amazon SageMaker Feature Store stores features in a collection called Feature Group. A Feature Group can be visualized as a table which has rows, with a unique identifier for each row where each column in the table is a feature. In principle, a Feature Group is composed of features and values per features.

Type: FeatureGroup object

Required: No

FeatureMetadata

The feature metadata used to search through the features.

Type: FeatureMetadata object

Required: No

HyperParameterTuningJob

The properties of a hyperparameter tuning job.

Type: HyperParameterTuningJobSearchEntity object

Required: No

Job

The properties of a job.

Type: Job object

Required: No

Model

A model displayed in the Amazon SageMaker Model Dashboard.

Type: ModelDashboardModel object

Required: No

ModelCard

An Amazon SageMaker Model Card that documents details about a machine learning model.

Type: ModelCard object

Required: No

ModelPackage

A container for your trained model that can be deployed for SageMaker inference. This can include inference code, artifacts, and metadata. The model package type can be one of the following.

- Versioned model: A part of a model package group in Model Registry.
- Unversioned model: Not part of a model package group and used in AWS Marketplace.

For more information, see `CreateModelPackage`.

Type: ModelPackage object

Required: No

ModelPackageGroup

A group of versioned models in the Model Registry.

Type: ModelPackageGroup object

Required: No

Pipeline

A SageMaker Model Building Pipeline instance.

Type: Pipeline object

Required: No

PipelineExecution

An execution of a pipeline.

Type: PipelineExecution object

Required: No

PipelineVersion

The version of the pipeline.

Type: PipelineVersion object

Required: No

Project

The properties of a project.

Type: Project object

Required: No

TrainingJob

The properties of a training job.

Type: TrainingJob object

Required: No

Trial

The properties of a trial.

Type: Trial object

Required: No

TrialComponent

The properties of a trial component.

Type: TrialComponent object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SecondaryStatusTransition

Service: Amazon SageMaker Service

An array element of `SecondaryStatusTransitions` for DescribeTrainingJob. It provides additional details about a status that the training job has transitioned through. A training job can be in one of several states, for example, starting, downloading, training, or uploading. Within each state, there are a number of intermediate states. For example, within the starting state, SageMaker could be starting the training job or launching the ML instances. These transitional states are referred to as the job's secondary status.

Contents

StartTime

A timestamp that shows when the training job transitioned to the current secondary status state.

Type: Timestamp

Required: Yes

Status

Contains a secondary status information from a training job.

Status might be one of the following secondary statuses:

InProgress

- `Starting` - Starting the training job.
- `Downloading` - An optional stage for algorithms that support `File` training input mode. It indicates that data is being downloaded to the ML storage volumes.
- `Training` - Training is in progress.
- `Uploading` - Training is complete and the model artifacts are being uploaded to the S3 location.

Completed

- `Completed` - The training job has completed.

Failed

- `Failed` - The training job has failed. The reason for the failure is returned in the `FailureReason` field of `DescribeTrainingJobResponse`.

Stopped

- `MaxRuntimeExceeded` - The job stopped because it exceeded the maximum allowed runtime.
- `Stopped` - The training job has stopped.

Stopping

- `Stopping` - Stopping the training job.

We no longer support the following secondary statuses:

- `LaunchingMLInstances`
- `PreparingTrainingStack`
- `DownloadingTrainingImage`

Type: String

Valid Values: `Starting | LaunchingMLInstances | PreparingTrainingStack | Downloading | DownloadingTrainingImage | Training | Uploading | Stopping | Stopped | MaxRuntimeExceeded | Completed | Failed | Interrupted | MaxWaitTimeExceeded | Updating | Restarting | Pending`

Required: Yes

EndTime

A timestamp that shows when the training job transitioned out of this secondary status state into another secondary status state or when the training job has ended.

Type: Timestamp

Required: No

StatusMessage

A detailed description of the progress within a secondary status.

SageMaker provides secondary statuses and status messages that apply to each of them:

Starting

- Starting the training job.
- Launching requested ML instances.
- Insufficient capacity error from EC2 while launching instances, retrying!

- Launched instance was unhealthy, replacing it!
- Preparing the instances for training.

Training

- Training image download completed. Training in progress.

> **Important**
>
> Status messages are subject to change. Therefore, we recommend not including them in code that programmatically initiates actions. For examples, don't use status messages in if statements.

To have an overview of your training job's progress, view `TrainingJobStatus` and `SecondaryStatus` in DescribeTrainingJob, and `StatusMessage` together. For example, at the start of a training job, you might see the following:

- `TrainingJobStatus` - InProgress
- `SecondaryStatus` - Training
- `StatusMessage` - Downloading the training image

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SelectedStep

Service: Amazon SageMaker Service

A step selected to run in selective execution mode.

Contents

StepName

The name of the pipeline step.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SelectiveExecutionConfig

Service: Amazon SageMaker Service

The selective execution configuration applied to the pipeline run.

Contents

SelectedSteps

A list of pipeline steps to run. All step(s) in all path(s) between two selected steps should be included.

Type: Array of SelectedStep objects

Array Members: Minimum number of 1 item. Maximum number of 50 items.

Required: Yes

SourcePipelineExecutionArn

The ARN from a reference execution of the current pipeline. Used to copy input collaterals needed for the selected steps to run. The execution status of the pipeline can be either `Failed` or `Success`.

This field is required if the steps you specify for `SelectedSteps` depend on output collaterals from any non-specified pipeline steps. For more information, see Selective Execution for Pipeline Steps.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SelectiveExecutionResult

Service: Amazon SageMaker Service

The ARN from an execution of the current pipeline.

Contents

SourcePipelineExecutionArn

The ARN from an execution of the current pipeline.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:pipeline\/.*\/execution\/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ServerlessJobConfig

Service: Amazon SageMaker Service

The configuration for the serverless training job.

Contents

BaseModelArn

The base model Amazon Resource Name (ARN) in SageMaker Public Hub. SageMaker always selects the latest version of the provided model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:[a-z0-9-\.]{1,63}:sagemaker:\w+(?:-\w+)+:(\d{12}|aws):hub-content\/)[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}\/Model\/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}(\/\d{1,4}.\d{1,4}.\d{1,4})?`

Required: Yes

JobType

The serverless training job type.

Type: String

Valid Values: `FineTuning | Evaluation`

Required: Yes

AcceptEula

Specifies agreement to the model end-user license agreement (EULA). The `AcceptEula` value must be explicitly defined as `True` in order to accept the EULA that this model requires. You are responsible for reviewing and complying with any applicable license terms and making sure they are acceptable for your use case before downloading or using a model. For more information, see End-user license agreements section for more details on accepting the EULA.

Type: Boolean

Required: No

CustomizationTechnique

The model customization technique.

Type: String

Valid Values: `SFT | DPO | RLVR | RLAIF`

Required: No

EvaluationType

The evaluation job type. Required when serverless job type is `Evaluation`.

Type: String

Valid Values: `LLMAJEvaluation | CustomScorerEvaluation | BenchmarkEvaluation`

Required: No

EvaluatorArn

The evaluator Amazon Resource Name (ARN) used as reward function or reward prompt.

Type: String

Pattern: `.*`

Required: No

Peft

The parameter-efficient fine-tuning configuration.

Type: String

Valid Values: `LORA`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ServiceCatalogProvisionedProductDetails

Service: Amazon SageMaker Service

Details of a provisioned service catalog product. For information about service catalog, see What is AWS Service Catalog.

Contents

ProvisionedProductId

The ID of the provisioned product.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[a-zA-Z0-9_\-]*`

Required: No

ProvisionedProductStatusMessage

The current status of the product.

- `AVAILABLE` - Stable state, ready to perform any operation. The most recent operation succeeded and completed.
- `UNDER_CHANGE` - Transitive state. Operations performed might not have valid results. Wait for an AVAILABLE status before performing operations.
- `TAINTED` - Stable state, ready to perform any operation. The stack has completed the requested operation but is not exactly what was requested. For example, a request to update to a new version failed and the stack rolled back to the current version.
- `ERROR` - An unexpected error occurred. The provisioned product exists but the stack is not running. For example, CloudFormation received a parameter value that was not valid and could not launch the stack.
- `PLAN_IN_PROGRESS` - Transitive state. The plan operations were performed to provision a new product, but resources have not yet been created. After reviewing the list of resources to be created, execute the plan. Wait for an AVAILABLE status before performing operations.

Type: String

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ServiceCatalogProvisioningDetails

Service: Amazon SageMaker Service

Details that you specify to provision a service catalog product. For information about service catalog, see What is AWS Service Catalog.

Contents

ProductId

The ID of the product to provision.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[a-zA-Z0-9_\-]*`

Required: Yes

PathId

The path identifier of the product. This value is optional if the product has a default path, and required if the product has more than one path.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[a-zA-Z0-9_\-]*`

Required: No

ProvisioningArtifactId

The ID of the provisioning artifact.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[a-zA-Z0-9_\-]*`

Required: No

ProvisioningParameters

A list of key value pairs that you specify when you provision a product.

Type: Array of ProvisioningParameter objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ServiceCatalogProvisioningUpdateDetails

Service: Amazon SageMaker Service

Details that you specify to provision a service catalog product. For information about service catalog, see What is AWS Service Catalog.

Contents

ProvisioningArtifactId

The ID of the provisioning artifact.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[a-zA-Z0-9_\-]*`

Required: No

ProvisioningParameters

A list of key value pairs that you specify when you provision a product.

Type: Array of ProvisioningParameter objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SessionChainingConfig

Service: Amazon SageMaker Service

Contains information about attribute-based access control (ABAC) for a training job. The session chaining configuration uses Amazon Security Token Service (STS) for your training job to request temporary, limited-privilege credentials to tenants. For more information, see Attribute-based access control (ABAC) for multi-tenancy training.

Contents

EnableSessionTagChaining

Set to `True` to allow SageMaker to extract session tags from a training job creation role and reuse these tags when assuming the training job execution role.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ShadowModeConfig

Service: Amazon SageMaker Service

The configuration of ShadowMode inference experiment type, which specifies a production variant to take all the inference requests, and a shadow variant to which Amazon SageMaker replicates a percentage of the inference requests. For the shadow variant it also specifies the percentage of requests that Amazon SageMaker replicates.

Contents

ShadowModelVariants

List of shadow variant configurations.

Type: Array of ShadowModelVariantConfig objects

Array Members: Fixed number of 1 item.

Required: Yes

SourceModelVariantName

The name of the production variant, which takes all the inference requests.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ShadowModelVariantConfig

Service: Amazon SageMaker Service

The name and sampling percentage of a shadow variant.

Contents

SamplingPercentage

The percentage of inference requests that Amazon SageMaker replicates from the production variant to the shadow variant.

Type: Integer

Valid Range: Maximum value of 100.

Required: Yes

ShadowModelVariantName

The name of the shadow variant.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SharingSettings

Service: Amazon SageMaker Service

Specifies options for sharing Amazon SageMaker AI Studio notebooks. These settings are specified as part of `DefaultUserSettings` when the `CreateDomain` API is called, and as part of `UserSettings` when the `CreateUserProfile` API is called. When `SharingSettings` is not specified, notebook sharing isn't allowed.

Contents

NotebookOutputOption

Whether to include the notebook cell output when sharing the notebook. The default is `Disabled`.

Type: String

Valid Values: `Allowed | Disabled`

Required: No

S3KmsKeyId

When `NotebookOutputOption` is `Allowed`, the AWS Key Management Service (KMS) encryption key ID used to encrypt the notebook cell output in the Amazon S3 bucket.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

S3OutputPath

When `NotebookOutputOption` is `Allowed`, the Amazon S3 bucket used to store the shared notebook snapshots.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ShuffleConfig

Service: Amazon SageMaker Service

A configuration for a shuffle option for input data in a channel. If you use `S3Prefix` for `S3DataType`, the results of the S3 key prefix matches are shuffled. If you use `ManifestFile`, the order of the S3 object references in the `ManifestFile` is shuffled. If you use `AugmentedManifestFile`, the order of the JSON lines in the `AugmentedManifestFile` is shuffled. The shuffling order is determined using the `Seed` value.

For Pipe input mode, when `ShuffleConfig` is specified shuffling is done at the start of every epoch. With large datasets, this ensures that the order of the training data is different for each epoch, and it helps reduce bias and possible overfitting. In a multi-node training job when `ShuffleConfig` is combined with `S3DataDistributionType` of `ShardedByS3Key`, the data is shuffled across nodes so that the content sent to a particular node on the first epoch might be sent to a different node on the second epoch.

Contents

Seed

Determines the shuffling order in `ShuffleConfig` value.

Type: Long

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SourceAlgorithm

Service: Amazon SageMaker Service

Specifies an algorithm that was used to create the model package. The algorithm must be either an algorithm resource in your SageMaker account or an algorithm in AWS Marketplace that you are subscribed to.

Contents

AlgorithmName

The name of an algorithm that was used to create the model package. The algorithm must be either an algorithm resource in your SageMaker account or an algorithm in AWS Marketplace that you are subscribed to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 170.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-]*\/)?([a-zA-Z0-9]([a-zA-Z0-9-]){0,62})(?<!-)`

Required: Yes

ModelDataETag

The ETag associated with Model Data URL.

Type: String

Required: No

ModelDataSource

Specifies the location of ML model data to deploy during endpoint creation.

Type: ModelDataSource object

Required: No

ModelDataUrl

The Amazon S3 path where the model artifacts, which result from model training, are stored. This path must point to a single `gzip` compressed tar archive (`.tar.gz` suffix).

> **Note**
>
> The model artifacts must be in an S3 bucket that is in the same AWS region as the algorithm.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SourceAlgorithmSpecification

Service: Amazon SageMaker Service

A list of algorithms that were used to create a model package.

Contents

SourceAlgorithms

A list of the algorithms that were used to create a model package.

Type: Array of SourceAlgorithm objects

Array Members: Fixed number of 1 item.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SourceIpConfig

Service: Amazon SageMaker Service

A list of IP address ranges (CIDRs). Used to create an allow list of IP addresses for a private workforce. Workers will only be able to log in to their worker portal from an IP address within this range. By default, a workforce isn't restricted to specific IP addresses.

Contents

Cidrs

A list of one to ten Classless Inter-Domain Routing (CIDR) values.

Maximum: Ten CIDR values

> **Note**
>
> The following Length Constraints apply to individual CIDR values in the CIDR value list.

Type: Array of strings

Length Constraints: Minimum length of 4. Maximum length of 64.

Pattern: `(^(([0-9]|[1-9][0-9]|1[0-9]{2}|2[0-4][0-9]|25[0-5])\.){3}([0-9]|[1-9][0-9]|1[0-9]{2}|2[0-4][0-9]|25[0-5])(\/(3[0-2]|[1-2][0-9]|[0-9]))$)|(^s*((([0-9A-Fa-f]{1,4}:){7}([0-9A-Fa-f]{1,4}|:))|(([0-9A-Fa-f]{1,4}:){6}(:[0-9A-Fa-f]{1,4}|((25[0-5]|2[0-4]d|1dd|[1-9]?d)(.(25[0-5]|2[0-4]d|1dd|[1-9]?d)){3})|:))|(([0-9A-Fa-f]{1,4}:){5}(((:[0-9A-Fa-f]{1,4}){1,2})|:((25[0-5]|2[0-4]d|1dd|[1-9]?d)(.(25[0-5]|2[0-4]d|1dd|[1-9]?d)){3})|:))|(([0-9A-Fa-f]{1,4}:){4}(((:[0-9A-Fa-f]{1,4}){1,3})|((:[0-9A-Fa-f]{1,4})?:((25[0-5]|2[0-4]d|1dd|[1-9]?d)(.(25[0-5]|2[0-4]d|1dd|[1-9]?d)){3}))|:))|(([0-9A-Fa-f]{1,4}:){3}(((:[0-9A-Fa-f]{1,4}){1,4})|((:[0-9A-Fa-f]{1,4}){0,2}:((25[0-5]|2[0-4]d|1dd|[1-9]?d)(.(25[0-5]|2[0-4]d|1dd|[1-9]?d)){3}))|:))|(([0-9A-Fa-f]{1,4}:){2}(((:[0-9A-Fa-f]{1,4}){1,5})|((:[0-9A-Fa-f]{1,4}){0,3}:((25[0-5]|2[0-4]d|1dd|[1-9]?d)(.(25[0-5]|2[0-4]d|1dd|[1-9]?d)){3}))|:))|(([0-9A-Fa-f]{1,4}:){1}(((:[0-9A-Fa-f]{1,4}){1,6})|((:[0-9A-Fa-f]{1,4}){0,4}:((25[0-5]|2[0-4]d|1dd|[1-9]?d)(.(25[0-5]|2[0-4]d|1dd|[1-9]?d)){3}))|:))|(:(((:[0-9A-Fa-f]`

```
{1,4}){1,7})|((:[0-9A-Fa-f]{1,4}){0,5}:((25[0-5]|2[0-4]d|1dd|[1-9]?d)(.
(25[0-5]|2[0-4]d|1dd|[1-9]?d)){3}))|:)))(%.+)?s*(\/(12[0-8]|1[0-1][0-9]|
[1-9][0-9]|[0-9]))$)
```

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpaceAppLifecycleManagement

Service: Amazon SageMaker Service

Settings that are used to configure and manage the lifecycle of Amazon SageMaker Studio applications in a space.

Contents

IdleSettings

Settings related to idle shutdown of Studio applications.

Type: SpaceIdleSettings object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpaceCodeEditorAppSettings

Service: Amazon SageMaker Service

The application settings for a Code Editor space.

Contents

AppLifecycleManagement

Settings that are used to configure and manage the lifecycle of CodeEditor applications in a space.

Type: SpaceAppLifecycleManagement object

Required: No

DefaultResourceSpec

Specifies the ARN's of a SageMaker AI image and SageMaker AI image version, and the instance type that the version runs on.

> **Note**
>
> When both `SageMakerImageVersionArn` and `SageMakerImageArn` are passed, `SageMakerImageVersionArn` is used. Any updates to `SageMakerImageArn` will not take effect if `SageMakerImageVersionArn` already exists in the `ResourceSpec` because `SageMakerImageVersionArn` always takes precedence. To clear the value set for `SageMakerImageVersionArn`, pass `None` as the value.

Type: ResourceSpec object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpaceDetails

Service: Amazon SageMaker Service

The space's details.

Contents

CreationTime

The creation time.

Type: Timestamp

Required: No

DomainId

The ID of the associated domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: No

LastModifiedTime

The last modified time.

Type: Timestamp

Required: No

OwnershipSettingsSummary

Specifies summary information about the ownership settings.

Type: OwnershipSettingsSummary object

Required: No

SpaceDisplayName

The name of the space that appears in the Studio UI.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `(?!\s*$).+`

Required: No

SpaceName

The name of the space.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

SpaceSettingsSummary

Specifies summary information about the space settings.

Type: SpaceSettingsSummary object

Required: No

SpaceSharingSettingsSummary

Specifies summary information about the space sharing settings.

Type: SpaceSharingSettingsSummary object

Required: No

Status

The status.

Type: String

Valid Values: `Deleting | Failed | InService | Pending | Updating | Update_Failed | Delete_Failed`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpaceIdleSettings

Service: Amazon SageMaker Service

Settings related to idle shutdown of Studio applications in a space.

Contents

IdleTimeoutInMinutes

The time that SageMaker waits after the application becomes idle before shutting it down.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 525600.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpaceJupyterLabAppSettings

Service: Amazon SageMaker Service

The settings for the JupyterLab application within a space.

Contents

AppLifecycleManagement

Settings that are used to configure and manage the lifecycle of JupyterLab applications in a space.

Type: SpaceAppLifecycleManagement object

Required: No

CodeRepositories

A list of Git repositories that SageMaker automatically displays to users for cloning in the JupyterLab application.

Type: Array of CodeRepository objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

DefaultResourceSpec

Specifies the ARN's of a SageMaker AI image and SageMaker AI image version, and the instance type that the version runs on.

> **Note**
>
> When both `SageMakerImageVersionArn` and `SageMakerImageArn` are passed, `SageMakerImageVersionArn` is used. Any updates to `SageMakerImageArn` will not take effect if `SageMakerImageVersionArn` already exists in the `ResourceSpec` because `SageMakerImageVersionArn` always takes precedence. To clear the value set for `SageMakerImageVersionArn`, pass `None` as the value.

Type: ResourceSpec object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpaceSettings

Service: Amazon SageMaker Service

A collection of space settings.

Contents

AppType

The type of app created within the space.

If using the UpdateSpace API, you can't change the app type of your space by specifying a different value for this field.

Type: String

Valid Values: `JupyterServer | KernelGateway | DetailedProfiler | TensorBoard | CodeEditor | JupyterLab | RStudioServerPro | RSessionGateway | Canvas`

Required: No

CodeEditorAppSettings

The Code Editor application settings.

Type: SpaceCodeEditorAppSettings object

Required: No

CustomFileSystems

A file system, created by you, that you assign to a space for an Amazon SageMaker AI Domain. Permitted users can access this file system in Amazon SageMaker AI Studio.

Type: Array of CustomFileSystem objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

JupyterLabAppSettings

The settings for the JupyterLab application.

Type: SpaceJupyterLabAppSettings object

Required: No

JupyterServerAppSettings

The JupyterServer app settings.

Type: JupyterServerAppSettings object

Required: No

KernelGatewayAppSettings

The KernelGateway app settings.

Type: KernelGatewayAppSettings object

Required: No

RemoteAccess

A setting that enables or disables remote access for a SageMaker space. When enabled, this allows you to connect to the remote space from your local IDE.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

SpaceManagedResources

If you enable this option, SageMaker AI creates the following resources on your behalf when you create the space:

- The user profile that possesses the space.
- The app that the space contains.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

SpaceStorageSettings

The storage settings for a space.

Type: SpaceStorageSettings object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpaceSettingsSummary

Service: Amazon SageMaker Service

Specifies summary information about the space settings.

Contents

AppType

The type of app created within the space.

Type: String

Valid Values: `JupyterServer | KernelGateway | DetailedProfiler | TensorBoard | CodeEditor | JupyterLab | RStudioServerPro | RSessionGateway | Canvas`

Required: No

RemoteAccess

A setting that enables or disables remote access for a SageMaker space. When enabled, this allows you to connect to the remote space from your local IDE.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

SpaceStorageSettings

The storage settings for a space.

Type: SpaceStorageSettings object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpaceSharingSettings

Service: Amazon SageMaker Service

A collection of space sharing settings.

Contents

SharingType

Specifies the sharing type of the space.

Type: String

Valid Values: `Private | Shared`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpaceSharingSettingsSummary

Service: Amazon SageMaker Service

Specifies summary information about the space sharing settings.

Contents

SharingType

Specifies the sharing type of the space.

Type: String

Valid Values: `Private | Shared`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpaceStorageSettings

Service: Amazon SageMaker Service

The storage settings for a space.

Contents

EbsStorageSettings

A collection of EBS storage settings for a space.

Type: EbsStorageSettings object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Stairs

Service: Amazon SageMaker Service

Defines the stairs traffic pattern for an Inference Recommender load test. This pattern type consists of multiple steps where the number of users increases at each step.

Specify either the stairs or phases traffic pattern.

Contents

DurationInSeconds

Defines how long each traffic step should be.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

NumberOfSteps

Specifies how many steps to perform during traffic.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

UsersPerStep

Specifies how many new users to spawn in each step.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 3.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StoppingCondition

Service: Amazon SageMaker Service

Specifies a limit to how long a job can run. When the job reaches the time limit, SageMaker ends the job. Use this API to cap costs.

To stop a training job, SageMaker sends the algorithm the `SIGTERM` signal, which delays job termination for 120 seconds. Algorithms can use this 120-second window to save the model artifacts, so the results of training are not lost.

The training algorithms provided by SageMaker automatically save the intermediate results of a model training job when possible. This attempt to save artifacts is only a best effort case as model might not be in a state from which it can be saved. For example, if training has just started, the model might not be ready to save. When saved, this intermediate data is a valid model artifact. You can use it to create a model with `CreateModel`.

> **Note**
>
> The Neural Topic Model (NTM) currently does not support saving intermediate model artifacts. When training NTMs, make sure that the maximum runtime is sufficient for the training job to complete.

Contents

MaxPendingTimeInSeconds

The maximum length of time, in seconds, that a training or compilation job can be pending before it is stopped.

> **Note**
>
> When working with training jobs that use capacity from training plans, not all `Pending` job states count against the `MaxPendingTimeInSeconds` limit. The following scenarios do not increment the `MaxPendingTimeInSeconds` counter:
>
> - The plan is in a `Scheduled` state: Jobs queued (in `Pending` status) before a plan's start date (waiting for scheduled start time)
> - Between capacity reservations: Jobs temporarily back to `Pending` status between two capacity reservation periods

`MaxPendingTimeInSeconds` only increments when jobs are actively waiting for capacity in an `Active` plan.

Type: Integer

Valid Range: Minimum value of 7200. Maximum value of 2419200.

Required: No

MaxRuntimeInSeconds

The maximum length of time, in seconds, that a training or compilation job can run before it is stopped.

For compilation jobs, if the job does not complete during this time, a `TimeOut` error is generated. We recommend starting with 900 seconds and increasing as necessary based on your model.

For all other jobs, if the job does not complete during this time, SageMaker ends the job. When `RetryStrategy` is specified in the job request, `MaxRuntimeInSeconds` specifies the maximum time for all of the attempts in total, not each individual attempt. The default value is 1 day. The maximum value is 28 days.

The maximum time that a `TrainingJob` can run in total, including any time spent publishing metrics or archiving and uploading models after it has been stopped, is 30 days.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

MaxWaitTimeInSeconds

The maximum length of time, in seconds, that a managed Spot training job has to complete. It is the amount of time spent waiting for Spot capacity plus the amount of time the job can run. It must be equal to or greater than `MaxRuntimeInSeconds`. If the job does not complete during this time, SageMaker ends the job.

When `RetryStrategy` is specified in the job request, `MaxWaitTimeInSeconds` specifies the maximum time for all of the attempts in total, not each individual attempt.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StudioLifecycleConfigDetails

Service: Amazon SageMaker Service

Details of the Amazon SageMaker AI Studio Lifecycle Configuration.

Contents

CreationTime

The creation time of the Amazon SageMaker AI Studio Lifecycle Configuration.

Type: Timestamp

Required: No

LastModifiedTime

This value is equivalent to CreationTime because Amazon SageMaker AI Studio Lifecycle Configurations are immutable.

Type: Timestamp

Required: No

StudioLifecycleConfigAppType

The App type to which the Lifecycle Configuration is attached.

Type: String

Valid Values: `JupyterServer | KernelGateway | CodeEditor | JupyterLab`

Required: No

StudioLifecycleConfigArn

The Amazon Resource Name (ARN) of the Lifecycle Configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:studio-lifecycle-config/.*|None)`

Required: No

StudioLifecycleConfigName

The name of the Amazon SageMaker AI Studio Lifecycle Configuration.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StudioWebPortalSettings

Service: Amazon SageMaker Service

Studio settings. If these settings are applied on a user level, they take priority over the settings applied on a domain level.

Contents

ExecutionRoleSessionNameMode

The execution role session name mode. If this value is set to `USER_IDENTITY`, the session name of the execution role corresponds to the user's identity. For IAM domains, the session name is the IAM session name used to generate the presigned URL. For IAM Identity Center domains, the session name is the username of the associated IAM Identity Center user. If this value is set to `STATIC` or is not set, the session name defaults to `SageMaker`.

Type: String

Valid Values: `STATIC | USER_IDENTITY`

Required: No

HiddenAppTypes

The Applications supported in Studio that are hidden from the Studio left navigation pane.

Type: Array of strings

Valid Values: `JupyterServer | KernelGateway | DetailedProfiler | TensorBoard | CodeEditor | JupyterLab | RStudioServerPro | RSessionGateway | Canvas`

Required: No

HiddenInstanceTypes

The instance types you are hiding from the Studio user interface.

Type: Array of strings

Valid Values: `system | ml.t3.micro | ml.t3.small | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.8xlarge | ml.m5.12xlarge | ml.m5.16xlarge | ml.m5.24xlarge | ml.m5d.large | ml.m5d.xlarge | ml.m5d.2xlarge | ml.m5d.4xlarge | ml.m5d.8xlarge | ml.m5d.12xlarge`

| ml.m5d.16xlarge | ml.m5d.24xlarge | ml.c5.large | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.12xlarge | ml.c5.18xlarge | ml.c5.24xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p3dn.24xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.r5.large | ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge | ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.12xlarge | ml.g6e.16xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.geospatial.interactive | ml.p4d.24xlarge | ml.p4de.24xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.p5.48xlarge | ml.p5en.48xlarge | ml.p6-b200.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.r7i.large | ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.m6id.large | ml.m6id.xlarge | ml.m6id.2xlarge | ml.m6id.4xlarge | ml.m6id.8xlarge | ml.m6id.12xlarge | ml.m6id.16xlarge | ml.m6id.24xlarge | ml.m6id.32xlarge | ml.c6id.large | ml.c6id.xlarge | ml.c6id.2xlarge | ml.c6id.4xlarge | ml.c6id.8xlarge | ml.c6id.12xlarge | ml.c6id.16xlarge | ml.c6id.24xlarge | ml.c6id.32xlarge | ml.r6id.large | ml.r6id.xlarge | ml.r6id.2xlarge | ml.r6id.4xlarge | ml.r6id.8xlarge | ml.r6id.12xlarge | ml.r6id.16xlarge | ml.r6id.24xlarge | ml.r6id.32xlarge | ml.p5.4xlarge

Required: No

HiddenMlTools

The machine learning tools that are hidden from the Studio left navigation pane.

Type: Array of strings

Valid Values: `DataWrangler | FeatureStore | EmrClusters | AutoMl | Experiments | Training | ModelEvaluation | Pipelines | Models | JumpStart | InferenceRecommender | Endpoints | Projects | InferenceOptimization | PerformanceEvaluation | LakeraGuard | Comet | DeepchecksLLMEvaluation | Fiddler | HyperPodClusters | RunningInstances | Datasets | Evaluators`

Required: No

HiddenSageMakerImageVersionAliases

The version aliases you are hiding from the Studio user interface.

Type: Array of HiddenSageMakerImage objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SubscribedWorkteam

Service: Amazon SageMaker Service

Describes a work team of a vendor that does the labelling job.

Contents

WorkteamArn

The Amazon Resource Name (ARN) of the vendor that you have subscribed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:workteam/.*`

Required: Yes

ListingId

Marketplace product listing ID.

Type: String

Required: No

MarketplaceDescription

The description of the vendor from the Amazon Marketplace.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `.+`

Required: No

MarketplaceTitle

The title of the service provided by the vendor in the Amazon Marketplace.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `.+`

Required: No

SellerName

The name of the vendor in the Amazon Marketplace.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SuggestionQuery

Service: Amazon SageMaker Service

Specified in the GetSearchSuggestions request. Limits the property names that are included in the response.

Contents

PropertyNameQuery

Defines a property name hint. Only property names that begin with the specified hint are included in the response.

Type: PropertyNameQuery object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TabularJobConfig

Service: Amazon SageMaker Service

The collection of settings used by an AutoML job V2 for the tabular problem type.

Contents

TargetAttributeName

The name of the target variable in supervised learning, usually represented by 'y'.

Type: String

Length Constraints: Minimum length of 1.

Required: Yes

CandidateGenerationConfig

The configuration information of how model candidates are generated.

Type: CandidateGenerationConfig object

Required: No

CompletionCriteria

How long a job is allowed to run, or how many candidates a job is allowed to generate.

Type: AutoMLJobCompletionCriteria object

Required: No

FeatureSpecificationS3Uri

A URL to the Amazon S3 data source containing selected features from the input data source to run an Autopilot job V2. You can input `FeatureAttributeNames` (optional) in JSON format as shown below:

```
{ "FeatureAttributeNames":["col1", "col2", ...] }.
```

You can also specify the data type of the feature (optional) in the format shown below:

```
{ "FeatureDataTypes":{"col1":"numeric", "col2":"categorical" ... } }
```

> **Note**
>
> These column keys may not include the target column.

In ensembling mode, Autopilot only supports the following data types: `numeric`, `categorical`, `text`, and `datetime`. In HPO mode, Autopilot can support `numeric`, `categorical`, `text`, `datetime`, and `sequence`.

If only `FeatureDataTypes` is provided, the column keys (`col1`, `col2`,..) should be a subset of the column names in the input data.

If both `FeatureDataTypes` and `FeatureAttributeNames` are provided, then the column keys should be a subset of the column names provided in `FeatureAttributeNames`.

The key name `FeatureAttributeNames` is fixed. The values listed in `["col1", "col2", ...]` are case sensitive and should be a list of strings containing unique values that are a subset of the column names in the input data. The list of columns provided must not include the target column.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

GenerateCandidateDefinitionsOnly

Generates possible candidates without training the models. A model candidate is a combination of data preprocessors, algorithms, and algorithm parameter settings.

Type: Boolean

Required: No

Mode

The method that Autopilot uses to train the data. You can either specify the mode manually or let Autopilot choose for you based on the dataset size by selecting `AUTO`. In `AUTO` mode, Autopilot chooses `ENSEMBLING` for datasets smaller than 100 MB, and `HYPERPARAMETER_TUNING` for larger ones.

The ENSEMBLING mode uses a multi-stack ensemble model to predict classification and regression tasks directly from your dataset. This machine learning mode combines several base models to produce an optimal predictive model. It then uses a stacking ensemble method to combine predictions from contributing members. A multi-stack ensemble model can provide better performance over a single model by combining the predictive capabilities of multiple models. See Autopilot algorithm support for a list of algorithms supported by ENSEMBLING mode.

The HYPERPARAMETER_TUNING (HPO) mode uses the best hyperparameters to train the best version of a model. HPO automatically selects an algorithm for the type of problem you want to solve. Then HPO finds the best hyperparameters according to your objective metric. See Autopilot algorithm support for a list of algorithms supported by HYPERPARAMETER_TUNING mode.

Type: String

Valid Values: AUTO | ENSEMBLING | HYPERPARAMETER_TUNING

Required: No

ProblemType

The type of supervised learning problem available for the model candidates of the AutoML job V2. For more information, see SageMaker Autopilot problem types.

> **Note**
>
> You must either specify the type of supervised learning problem in ProblemType and provide the AutoMLJobObjective metric, or none at all.

Type: String

Valid Values: BinaryClassification | MulticlassClassification | Regression

Required: No

SampleWeightAttributeName

If specified, this column name indicates which column of the dataset should be treated as sample weights for use by the objective metric during the training, evaluation, and the selection

of the best model. This column is not considered as a predictive feature. For more information on Autopilot metrics, see Metrics and validation.

Sample weights should be numeric, non-negative, with larger values indicating which rows are more important than others. Data points that have invalid or no weight value are excluded.

Support for sample weights is available in Ensembling mode only.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9_-]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TabularResolvedAttributes

Service: Amazon SageMaker Service

The resolved attributes specific to the tabular problem type.

Contents

ProblemType

The type of supervised learning problem available for the model candidates of the AutoML job V2 (Binary Classification, Multiclass Classification, Regression). For more information, see SageMaker Autopilot problem types.

Type: String

Valid Values: `BinaryClassification | MulticlassClassification | Regression`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Tag

Service: Amazon SageMaker Service

A tag object that consists of a key and an optional value, used to manage metadata for SageMaker AWS resources.

You can add tags to notebook instances, training jobs, hyperparameter tuning jobs, batch transform jobs, models, labeling jobs, work teams, endpoint configurations, and endpoints. For more information on adding tags to SageMaker resources, see AddTags.

For more information on adding metadata to your AWS resources with tagging, see Tagging AWS resources. For advice on best practices for managing AWS resources with tagging, see Tagging Best Practices: Implement an Effective AWS Resource Tagging Strategy.

Contents

Key

The tag key. Tag keys must be unique per resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `([\p{L}\p{Z}\p{N}_.:/=+\-@]*)`

Required: Yes

Value

The tag value.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `([\p{L}\p{Z}\p{N}_.:/=+\-@]*)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TargetPlatform

Service: Amazon SageMaker Service

Contains information about a target platform that you want your model to run on, such as OS, architecture, and accelerators. It is an alternative of `TargetDevice`.

Contents

Arch

Specifies a target platform architecture.

- `X86_64`: 64-bit version of the x86 instruction set.
- `X86`: 32-bit version of the x86 instruction set.
- `ARM64`: ARMv8 64-bit CPU.
- `ARM_EABIHF`: ARMv7 32-bit, Hard Float.
- `ARM_EABI`: ARMv7 32-bit, Soft Float. Used by Android 32-bit ARM platform.

Type: String

Valid Values: `X86_64 | X86 | ARM64 | ARM_EABI | ARM_EABIHF`

Required: Yes

Os

Specifies a target platform OS.

- `LINUX`: Linux-based operating systems.
- `ANDROID`: Android operating systems. Android API level can be specified using the `ANDROID_PLATFORM` compiler option. For example, `"CompilerOptions": {'ANDROID_PLATFORM': 28}`

Type: String

Valid Values: `ANDROID | LINUX`

Required: Yes

Accelerator

Specifies a target platform accelerator (optional).

- `NVIDIA`: Nvidia graphics processing unit. It also requires `gpu-code`, `trt-ver`, `cuda-ver` compiler options
- `MALI`: ARM Mali graphics processor
- `INTEL_GRAPHICS`: Integrated Intel graphics

Type: String

Valid Values: `INTEL_GRAPHICS | MALI | NVIDIA | NNA`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TargetTrackingScalingPolicyConfiguration

Service: Amazon SageMaker Service

A target tracking scaling policy. Includes support for predefined or customized metrics.

When using the PutScalingPolicy API, this parameter is required when you are creating a policy with the policy type `TargetTrackingScaling`.

Contents

MetricSpecification

An object containing information about a metric.

Type: MetricSpecification object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

TargetValue

The recommended target value to specify for the metric when creating a scaling policy.

Type: Double

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TemplateProviderDetail

Service: Amazon SageMaker Service

Details about a template provider configuration and associated provisioning information.

Contents

CfnTemplateProviderDetail

Details about a CloudFormation template provider configuration and associated provisioning information.

Type: CfnTemplateProviderDetail object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TensorBoardAppSettings

Service: Amazon SageMaker Service

The TensorBoard app settings.

Contents

DefaultResourceSpec

The default instance type and the Amazon Resource Name (ARN) of the SageMaker AI image created on the instance.

Type: ResourceSpec object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TensorBoardOutputConfig

Service: Amazon SageMaker Service

Configuration of storage locations for the Amazon SageMaker Debugger TensorBoard output data.

Contents

S3OutputPath

Path to Amazon S3 storage location for TensorBoard output.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

LocalPath

Path to local storage location for tensorBoard output. Defaults to `/opt/ml/output/tensorboard`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4096.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TextClassificationJobConfig

Service: Amazon SageMaker Service

The collection of settings used by an AutoML job V2 for the text classification problem type.

Contents

ContentColumn

The name of the column used to provide the sentences to be classified. It should not be the same as the target column.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: Yes

TargetLabelColumn

The name of the column used to provide the class labels. It should not be same as the content column.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: Yes

CompletionCriteria

How long a job is allowed to run, or how many candidates a job is allowed to generate.

Type: AutoMLJobCompletionCriteria object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

TextGenerationJobConfig

Service: Amazon SageMaker Service

The collection of settings used by an AutoML job V2 for the text generation problem type.

> **Note**
>
> The text generation models that support fine-tuning in Autopilot are currently accessible exclusively in regions supported by Canvas. Refer to the documentation of Canvas for the full list of its supported Regions.

Contents

BaseModelName

The name of the base model to fine-tune. Autopilot supports fine-tuning a variety of large language models. For information on the list of supported models, see Text generation models supporting fine-tuning in Autopilot. If no `BaseModelName` is provided, the default model used is **Falcon7BInstruct**.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

CompletionCriteria

How long a fine-tuning job is allowed to run. For `TextGenerationJobConfig` problem types, the `MaxRuntimePerTrainingJobInSeconds` attribute of `AutoMLJobCompletionCriteria` defaults to 72h (259200s).

Type: AutoMLJobCompletionCriteria object

Required: No

ModelAccessConfig

The access configuration file to control access to the ML model. You can explicitly accept the model end-user license agreement (EULA) within the `ModelAccessConfig`.

- If you are a Jumpstart user, see the End-user license agreements section for more details on accepting the EULA.
- If you are an AutoML user, see the *Optional Parameters* section of *Create an AutoML job to fine-tune text generation models using the API* for details on How to set the EULA acceptance when fine-tuning a model using the AutoML API.

Type: ModelAccessConfig object

Required: No

TextGenerationHyperParameters

The hyperparameters used to configure and optimize the learning process of the base model. You can set any combination of the following hyperparameters for all base models. For more information on each supported hyperparameter, see Optimize the learning process of your text generation models with hyperparameters.

- `"epochCount"`: The number of times the model goes through the entire training dataset. Its value should be a string containing an integer value within the range of "1" to "10".
- `"batchSize"`: The number of data samples used in each iteration of training. Its value should be a string containing an integer value within the range of "1" to "64".
- `"learningRate"`: The step size at which a model's parameters are updated during training. Its value should be a string containing a floating-point value within the range of "0" to "1".
- `"learningRateWarmupSteps"`: The number of training steps during which the learning rate gradually increases before reaching its target or maximum value. Its value should be a string containing an integer value within the range of "0" to "250".

Here is an example where all four hyperparameters are configured.

```
{ "epochCount":"5", "learningRate":"0.5", "batchSize": "32",
"learningRateWarmupSteps": "10" }
```

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 30 items.

Key Length Constraints: Minimum length of 0. Maximum length of 32.

Key Pattern: `[a-zA-Z0-9._-]+`

Value Length Constraints: Minimum length of 0. Maximum length of 16.

Value Pattern: `[a-zA-Z0-9._-]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TextGenerationResolvedAttributes

Service: Amazon SageMaker Service

The resolved attributes specific to the text generation problem type.

Contents

BaseModelName

The name of the base model to fine-tune.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ThroughputConfig

Service: Amazon SageMaker Service

Used to set feature group throughput configuration. There are two modes: `ON_DEMAND` and `PROVISIONED`. With on-demand mode, you are charged for data reads and writes that your application performs on your feature group. You do not need to specify read and write throughput because Feature Store accommodates your workloads as they ramp up and down. You can switch a feature group to on-demand only once in a 24 hour period. With provisioned throughput mode, you specify the read and write capacity per second that you expect your application to require, and you are billed based on those limits. Exceeding provisioned throughput will result in your requests being throttled.

Note: `PROVISIONED` throughput mode is supported only for feature groups that are offline-only, or use the `Standard` tier online store.

Contents

ThroughputMode

The mode used for your feature group throughput: `ON_DEMAND` or `PROVISIONED`.

Type: String

Valid Values: `OnDemand | Provisioned`

Required: Yes

ProvisionedReadCapacityUnits

For provisioned feature groups with online store enabled, this indicates the read throughput you are billed for and can consume without throttling.

This field is not applicable for on-demand feature groups.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

ProvisionedWriteCapacityUnits

For provisioned feature groups, this indicates the write throughput you are billed for and can consume without throttling.

This field is not applicable for on-demand feature groups.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ThroughputConfigDescription

Service: Amazon SageMaker Service

Active throughput configuration of the feature group. There are two modes: `ON_DEMAND` and `PROVISIONED`. With on-demand mode, you are charged for data reads and writes that your application performs on your feature group. You do not need to specify read and write throughput because Feature Store accommodates your workloads as they ramp up and down. You can switch a feature group to on-demand only once in a 24 hour period. With provisioned throughput mode, you specify the read and write capacity per second that you expect your application to require, and you are billed based on those limits. Exceeding provisioned throughput will result in your requests being throttled.

Note: `PROVISIONED` throughput mode is supported only for feature groups that are offline-only, or use the `Standard` tier online store.

Contents

ThroughputMode

The mode used for your feature group throughput: `ON_DEMAND` or `PROVISIONED`.

Type: String

Valid Values: `OnDemand | Provisioned`

Required: Yes

ProvisionedReadCapacityUnits

For provisioned feature groups with online store enabled, this indicates the read throughput you are billed for and can consume without throttling.

This field is not applicable for on-demand feature groups.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

ProvisionedWriteCapacityUnits

For provisioned feature groups, this indicates the write throughput you are billed for and can consume without throttling.

This field is not applicable for on-demand feature groups.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ThroughputConfigUpdate

Service: Amazon SageMaker Service

The new throughput configuration for the feature group. You can switch between on-demand and provisioned modes or update the read / write capacity of provisioned feature groups. You can switch a feature group to on-demand only once in a 24 hour period.

Contents

ProvisionedReadCapacityUnits

For provisioned feature groups with online store enabled, this indicates the read throughput you are billed for and can consume without throttling.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

ProvisionedWriteCapacityUnits

For provisioned feature groups, this indicates the write throughput you are billed for and can consume without throttling.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 10000000.

Required: No

ThroughputMode

Target throughput mode of the feature group. Throughput update is an asynchronous operation, and the outcome should be monitored by polling `LastUpdateStatus` field in `DescribeFeatureGroup` response. You cannot update a feature group's throughput while another update is in progress.

Type: String

Valid Values: `OnDemand | Provisioned`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TimeSeriesConfig

Service: Amazon SageMaker Service

The collection of components that defines the time-series.

Contents

ItemIdentifierAttributeName

The name of the column that represents the set of item identifiers for which you want to predict the target value.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: Yes

TargetAttributeName

The name of the column representing the target variable that you want to predict for each item in your dataset. The data type of the target variable must be numerical.

Type: String

Length Constraints: Minimum length of 1.

Required: Yes

TimestampAttributeName

The name of the column indicating a point in time at which the target value of a given item is recorded.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: Yes

GroupingAttributeNames

A set of columns names that can be grouped with the item identifier column to create a composite key for which a target value is predicted.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TimeSeriesForecastingJobConfig

Service: Amazon SageMaker Service

The collection of settings used by an AutoML job V2 for the time-series forecasting problem type.

Contents

ForecastFrequency

The frequency of predictions in a forecast.

Valid intervals are an integer followed by Y (Year), M (Month), W (Week), D (Day), H (Hour), and min (Minute). For example, `1D` indicates every day and `15min` indicates every 15 minutes. The value of a frequency must not overlap with the next larger frequency. For example, you must use a frequency of `1H` instead of `60min`.

The valid values for each frequency are the following:

- Minute - 1-59
- Hour - 1-23
- Day - 1-6
- Week - 1-4
- Month - 1-11
- Year - 1

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `1Y|Y|([1-9]|1[0-1])M|M|[1-4]W|W|[1-6]D|D|([1-9]|1[0-9]|2[0-3])H|H|([1-9]|[1-5][0-9])min`

Required: Yes

ForecastHorizon

The number of time-steps that the model predicts. The forecast horizon is also called the prediction length. The maximum forecast horizon is the lesser of 500 time-steps or 1/4 of the time-steps in the dataset.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

TimeSeriesConfig

The collection of components that defines the time-series.

Type: TimeSeriesConfig object

Required: Yes

CandidateGenerationConfig

Stores the configuration information for how model candidates are generated using an AutoML job V2.

Type: CandidateGenerationConfig object

Required: No

CompletionCriteria

How long a job is allowed to run, or how many candidates a job is allowed to generate.

Type: AutoMLJobCompletionCriteria object

Required: No

FeatureSpecificationS3Uri

A URL to the Amazon S3 data source containing additional selected features that complement the target, itemID, timestamp, and grouped columns set in `TimeSeriesConfig`. When not provided, the AutoML job V2 includes all the columns from the original dataset that are not already declared in `TimeSeriesConfig`. If provided, the AutoML job V2 only considers these additional columns as a complement to the ones declared in `TimeSeriesConfig`.

You can input `FeatureAttributeNames` (optional) in JSON format as shown below:

```
{ "FeatureAttributeNames":["col1", "col2", ...] }.
```

You can also specify the data type of the feature (optional) in the format shown below:

```
{ "FeatureDataTypes":{"col1":"numeric", "col2":"categorical" ... } }
```

Autopilot supports the following data types: `numeric`, `categorical`, `text`, and `datetime`.

> **Note**
>
> These column keys must not include any column set in `TimeSeriesConfig`.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

ForecastQuantiles

The quantiles used to train the model for forecasts at a specified quantile. You can specify quantiles from `0.01` (p1) to `0.99` (p99), by increments of 0.01 or higher. Up to five forecast quantiles can be specified. When `ForecastQuantiles` is not provided, the AutoML job uses the quantiles p10, p50, and p90 as default.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 2. Maximum length of 4.

Pattern: `(^p[1-9]\d?$)`

Required: No

HolidayConfig

The collection of holiday featurization attributes used to incorporate national holiday information into your forecasting model.

Type: Array of HolidayConfigAttributes objects

Array Members: Fixed number of 1 item.

Required: No

Transformations

The transformations modifying specific attributes of the time-series, such as filling strategies for missing values.

Type: TimeSeriesTransformations object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TimeSeriesForecastingSettings

Service: Amazon SageMaker Service

Time series forecast settings for the SageMaker Canvas application.

Contents

AmazonForecastRoleArn

The IAM role that Canvas passes to Amazon Forecast for time series forecasting. By default, Canvas uses the execution role specified in the `UserProfile` that launches the Canvas application. If an execution role is not specified in the `UserProfile`, Canvas uses the execution role specified in the Domain that owns the `UserProfile`. To allow time series forecasting, this IAM role should have the AmazonSageMakerCanvasForecastAccess policy attached and `forecast.amazonaws.com` added in the trust relationship as a service principal.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

Status

Describes whether time series forecasting is enabled or disabled in the Canvas application.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

TimeSeriesTransformations

Service: Amazon SageMaker Service

Transformations allowed on the dataset. Supported transformations are `Filling` and `Aggregation`. `Filling` specifies how to add values to missing values in the dataset. `Aggregation` defines how to aggregate data that does not align with forecast frequency.

Contents

Aggregation

A key value pair defining the aggregation method for a column, where the key is the column name and the value is the aggregation method.

The supported aggregation methods are `sum` (default), `avg`, `first`, `min`, `max`.

> **Note**
>
> Aggregation is only supported for the target column.

Type: String to string map

Map Entries: Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Valid Values: `sum | avg | first | min | max`

Required: No

Filling

A key value pair defining the filling method for a column, where the key is the column name and the value is an object which defines the filling logic. You can specify multiple filling methods for a single column.

The supported filling methods and their corresponding options are:

- `frontfill`: none (Supported only for target column)
- `middlefill`: `zero`, `value`, `median`, `mean`, `min`, `max`
- `backfill`: `zero`, `value`, `median`, `mean`, `min`, `max`

- `futurefill`: `zero`, `value`, `median`, `mean`, `min`, `max`

  To set a filling method to a specific value, set the fill parameter to the chosen filling method value (for example `"backfill" : "value"`), and define the filling value in an additional parameter prefixed with "_value". For example, to set `backfill` to a value of 2, you must include two parameters: `"backfill": "value"` and `"backfill_value":"2"`.

  Type: String to string to string map map

  Map Entries: Maximum number of 50 items.

  Key Length Constraints: Minimum length of 1. Maximum length of 256.

  Map Entries: Maximum number of 6 items.

  Valid Keys: `frontfill | middlefill | backfill | futurefill | frontfill_value | middlefill_value | backfill_value | futurefill_value`

  Value Length Constraints: Minimum length of 1. Maximum length of 256.

  Value Pattern: `[a-zA-Z0-9\_\-]+`

  Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TotalHits

Service: Amazon SageMaker Service

Represents the total number of matching results and indicates how accurate that count is.

The `Value` field provides the count, which may be exact or estimated. The `Relation` field indicates whether it's an exact figure or a lower bound. This helps understand the full scope of search results, especially when dealing with large result sets.

Contents

Relation

Indicates the relationship between the returned `Value` and the actual total number of matching results. Possible values are:

- `EqualTo`: The `Value` is the exact count of matching results.
- `GreaterThanOrEqualTo`: The `Value` is a lower bound of the actual count of matching results.

Type: String

Valid Values: `EqualTo | GreaterThanOrEqualTo`

Required: No

Value

The total number of matching results. This value may be exact or an estimate, depending on the `Relation` field.

Type: Long

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrackingServerSummary

Service: Amazon SageMaker Service

The summary of the tracking server to list.

Contents

CreationTime

The creation time of a listed tracking server.

Type: Timestamp

Required: No

IsActive

The activity status of a listed tracking server.

Type: String

Valid Values: `Active | Inactive`

Required: No

LastModifiedTime

The last modified time of a listed tracking server.

Type: Timestamp

Required: No

MlflowVersion

The MLflow version used for a listed tracking server.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 16.

Pattern: `[0-9]*.[0-9]*.[0-9]*`

Required: No

TrackingServerArn

The ARN of a listed tracking server.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:mlflow-tracking-server/.*`

Required: No

TrackingServerName

The name of a listed tracking server.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,255}`

Required: No

TrackingServerStatus

The creation status of a listed tracking server.

Type: String

Valid Values: `Creating | Created | CreateFailed | Updating | Updated | UpdateFailed | Deleting | DeleteFailed | Stopping | Stopped | StopFailed | Starting | Started | StartFailed | MaintenanceInProgress | MaintenanceComplete | MaintenanceFailed`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrafficPattern

Service: Amazon SageMaker Service

Defines the traffic pattern of the load test.

Contents

Phases

Defines the phases traffic specification.

Type: Array of Phase objects

Array Members: Minimum number of 1 item.

Required: No

Stairs

Defines the stairs traffic pattern.

Type: Stairs object

Required: No

TrafficType

Defines the traffic patterns. Choose either `PHASES` or `STAIRS`.

Type: String

Valid Values: `PHASES | STAIRS`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrafficRoutingConfig

Service: Amazon SageMaker Service

Defines the traffic routing strategy during an endpoint deployment to shift traffic from the old fleet to the new fleet.

Contents

Type

Traffic routing strategy type.

- `ALL_AT_ONCE`: Endpoint traffic shifts to the new fleet in a single step.
- `CANARY`: Endpoint traffic shifts to the new fleet in two steps. The first step is the canary, which is a small portion of the traffic. The second step is the remainder of the traffic.
- `LINEAR`: Endpoint traffic shifts to the new fleet in n steps of a configurable size.

Type: String

Valid Values: `ALL_AT_ONCE | CANARY | LINEAR`

Required: Yes

WaitIntervalInSeconds

The waiting time (in seconds) between incremental steps to turn on traffic on the new endpoint fleet.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 3600.

Required: Yes

CanarySize

Batch size for the first step to turn on traffic on the new endpoint fleet. `Value` must be less than or equal to 50% of the variant's total instance count.

Type: CapacitySize object

Required: No

LinearStepSize

Batch size for each step to turn on traffic on the new endpoint fleet. `Value` must be 10-50% of the variant's total instance count.

Type: CapacitySize object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingImageConfig

Service: Amazon SageMaker Service

The configuration to use an image from a private Docker registry for a training job.

Contents

TrainingRepositoryAccessMode

The method that your training job will use to gain access to the images in your private Docker registry. For access to an image in a private Docker registry, set to `Vpc`.

Type: String

Valid Values: `Platform | Vpc`

Required: Yes

TrainingRepositoryAuthConfig

An object containing authentication information for a private Docker registry containing your training images.

Type: TrainingRepositoryAuthConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingJob

Service: Amazon SageMaker Service

Contains information about a training job.

Contents

AlgorithmSpecification

Information about the algorithm used for training, and algorithm metadata.

Type: AlgorithmSpecification object

Required: No

AutoMLJobArn

The Amazon Resource Name (ARN) of the job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

Required: No

BillableTimeInSeconds

The billable time in seconds.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

CheckpointConfig

Contains information about the output location for managed spot training checkpoint data.

Type: CheckpointConfig object

Required: No

CreationTime

A timestamp that indicates when the training job was created.

Type: Timestamp

Required: No

DebugHookConfig

Configuration information for the Amazon SageMaker Debugger hook parameters, metric and tensor collections, and storage paths. To learn more about how to configure the `DebugHookConfig` parameter, see Use the SageMaker and Debugger Configuration API Operations to Create, Update, and Debug Your Training Job.

Type: DebugHookConfig object

Required: No

DebugRuleConfigurations

Information about the debug rule configuration.

Type: Array of DebugRuleConfiguration objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

DebugRuleEvaluationStatuses

Information about the evaluation status of the rules for the training job.

Type: Array of DebugRuleEvaluationStatus objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

EnableInterContainerTrafficEncryption

To encrypt all communications between ML compute instances in distributed training, choose `True`. Encryption provides greater security for distributed training, but training might take longer. How long it takes depends on the amount of communication between compute instances, especially if you use a deep learning algorithm in distributed training.

Type: Boolean

Required: No

EnableManagedSpotTraining

When true, enables managed spot training using Amazon EC2 Spot instances to run training jobs instead of on-demand instances. For more information, see Managed Spot Training.

Type: Boolean

Required: No

EnableNetworkIsolation

If the `TrainingJob` was created with network isolation, the value is set to `true`. If network isolation is enabled, nodes can't communicate beyond the VPC they run in.

Type: Boolean

Required: No

Environment

The environment variables to set in the Docker container.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 512.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]*`

Value Length Constraints: Minimum length of 0. Maximum length of 512.

Value Pattern: `[\S\s]*`

Required: No

ExperimentConfig

Associates a SageMaker job as a trial component with an experiment and trial. Specified when you call the following APIs:

- CreateProcessingJob

- CreateTrainingJob
- CreateTransformJob

Type: ExperimentConfig object

Required: No

FailureReason

If the training job failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

FinalMetricDataList

A list of final metric values that are set when the training job completes. Used only if the training job was configured to use metrics.

Type: Array of MetricData objects

Array Members: Minimum number of 0 items. Maximum number of 40 items.

Required: No

HyperParameters

Algorithm-specific parameters.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Required: No

InputDataConfig

An array of `Channel` objects that describes each data input channel.

Your input must be in the same AWS region as your training job.

Type: Array of Channel objects

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Required: No

LabelingJobArn

The Amazon Resource Name (ARN) of the labeling job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:labeling-job/.*`

Required: No

LastModifiedTime

A timestamp that indicates when the status of the training job was last modified.

Type: Timestamp

Required: No

ModelArtifacts

Information about the Amazon S3 location that is configured for storing model artifacts.

Type: ModelArtifacts object

Required: No

ModelPackageConfig

The model package configuration.

Type: ModelPackageConfig object

Required: No

OutputDataConfig

The S3 path where model artifacts that you configured when creating the job are stored. SageMaker creates subfolders for model artifacts.

Type: OutputDataConfig object

Required: No

OutputModelPackageArn

The output model package Amazon Resource Name (ARN) that contains model weights or checkpoint.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: No

ProfilerConfig

Configuration information for Amazon SageMaker Debugger system monitoring, framework profiling, and storage paths.

Type: ProfilerConfig object

Required: No

ResourceConfig

Resources, including ML compute instances and ML storage volumes, that are configured for model training.

Type: ResourceConfig object

Required: No

RetryStrategy

The number of times to retry the job when the job fails due to an `InternalServerError`.

Type: RetryStrategy object

Required: No

RoleArn

The AWS Identity and Access Management (IAM) role configured for the training job.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

SecondaryStatus

Provides detailed information about the state of the training job. For detailed information about the secondary status of the training job, see `StatusMessage` under SecondaryStatusTransition.

SageMaker provides primary statuses and secondary statuses that apply to each of them:

InProgress

- `Starting` - Starting the training job.
- `Downloading` - An optional stage for algorithms that support `File` training input mode. It indicates that data is being downloaded to the ML storage volumes.
- `Training` - Training is in progress.
- `Uploading` - Training is complete and the model artifacts are being uploaded to the S3 location.

Completed

- `Completed` - The training job has completed.

Failed

- `Failed` - The training job has failed. The reason for the failure is returned in the `FailureReason` field of `DescribeTrainingJobResponse`.

Stopped

- `MaxRuntimeExceeded` - The job stopped because it exceeded the maximum allowed runtime.
- `Stopped` - The training job has stopped.

Stopping

- `Stopping` - Stopping the training job.

> **Important**
>
> Valid values for `SecondaryStatus` are subject to change.

We no longer support the following secondary statuses:

- `LaunchingMLInstances`
- `PreparingTrainingStack`
- `DownloadingTrainingImage`

Type: String

Valid Values: `Starting | LaunchingMLInstances | PreparingTrainingStack | Downloading | DownloadingTrainingImage | Training | Uploading | Stopping | Stopped | MaxRuntimeExceeded | Completed | Failed | Interrupted | MaxWaitTimeExceeded | Updating | Restarting | Pending`

Required: No

SecondaryStatusTransitions

A history of all of the secondary statuses that the training job has transitioned through.

Type: Array of SecondaryStatusTransition objects

Required: No

StoppingCondition

Specifies a limit to how long a model training job can run. It also specifies how long a managed Spot training job has to complete. When the job reaches the time limit, SageMaker ends the training job. Use this API to cap model training costs.

To stop a job, SageMaker sends the algorithm the `SIGTERM` signal, which delays job termination for 120 seconds. Algorithms can use this 120-second window to save the model artifacts, so the results of training are not lost.

Type: StoppingCondition object

Required: No

Tags

An array of key-value pairs. You can use tags to categorize your AWS resources in different ways, for example, by purpose, owner, or environment. For more information, see Tagging AWS Resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TensorBoardOutputConfig

Configuration of storage locations for the Amazon SageMaker Debugger TensorBoard output data.

Type: TensorBoardOutputConfig object

Required: No

TrainingEndTime

Indicates the time when the training job ends on training instances. You are billed for the time interval between the value of `TrainingStartTime` and this time. For successful jobs and stopped jobs, this is the time after model artifacts are uploaded. For failed jobs, this is the time when SageMaker detects a job failure.

Type: Timestamp

Required: No

TrainingJobArn

The Amazon Resource Name (ARN) of the training job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

TrainingJobName

The name of the training job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

TrainingJobStatus

The status of the training job.

Training job statuses are:

- `InProgress` - The training is in progress.
- `Completed` - The training job has completed.
- `Failed` - The training job has failed. To see the reason for the failure, see the `FailureReason` field in the response to a `DescribeTrainingJobResponse` call.
- `Stopping` - The training job is stopping.
- `Stopped` - The training job has stopped.

For more detailed information, see `SecondaryStatus`.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting`

Required: No

TrainingStartTime

Indicates the time when the training job starts on training instances. You are billed for the time interval between this time and the value of `TrainingEndTime`. The start time in CloudWatch Logs might be later than this time. The difference is due to the time it takes to download the training data and to the size of the training container.

Type: Timestamp

Required: No

TrainingTimeInSeconds

The training time in seconds.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

TuningJobArn

The Amazon Resource Name (ARN) of the associated hyperparameter tuning job if the training job was launched by a hyperparameter tuning job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:hyper-parameter-tuning-job/.*`

Required: No

VpcConfig

A VpcConfig object that specifies the VPC that this training job has access to. For more information, see Protect Training Jobs by Using an Amazon Virtual Private Cloud.

Type: VpcConfig object

Required: No

WarmPoolStatus

The status of the warm pool associated with the training job.

Type: WarmPoolStatus object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingJobDefinition

Service: Amazon SageMaker Service

Defines the input needed to run a training job using the algorithm.

Contents

InputDataConfig

An array of `Channel` objects, each of which specifies an input source.

Type: Array of Channel objects

Array Members: Minimum number of 1 item. Maximum number of 20 items.

Required: Yes

OutputDataConfig

the path to the S3 bucket where you want to store model artifacts. SageMaker creates subfolders for the artifacts.

Type: OutputDataConfig object

Required: Yes

ResourceConfig

The resources, including the ML compute instances and ML storage volumes, to use for model training.

Type: ResourceConfig object

Required: Yes

StoppingCondition

Specifies a limit to how long a model training job can run. It also specifies how long a managed Spot training job has to complete. When the job reaches the time limit, SageMaker ends the training job. Use this API to cap model training costs.

To stop a job, SageMaker sends the algorithm the SIGTERM signal, which delays job termination for 120 seconds. Algorithms can use this 120-second window to save the model artifacts.

Type: StoppingCondition object

Required: Yes

TrainingInputMode

The training input mode that the algorithm supports. For more information about input modes, see Algorithms.

Pipe mode

If an algorithm supports `Pipe` mode, Amazon SageMaker streams data directly from Amazon S3 to the container.

File mode

If an algorithm supports `File` mode, SageMaker downloads the training data from S3 to the provisioned ML storage volume, and mounts the directory to the Docker volume for the training container.

You must provision the ML storage volume with sufficient capacity to accommodate the data downloaded from S3. In addition to the training data, the ML storage volume also stores the output model. The algorithm container uses the ML storage volume to also store intermediate information, if any.

For distributed algorithms, training data is distributed uniformly. Your training duration is predictable if the input data objects sizes are approximately the same. SageMaker does not split the files any further for model training. If the object sizes are skewed, training won't be optimal as the data distribution is also skewed when one host in a training cluster is overloaded, thus becoming a bottleneck in training.

FastFile mode

If an algorithm supports `FastFile` mode, SageMaker streams data directly from S3 to the container with no code changes, and provides file system access to the data. Users can author their training script to interact with these files as if they were stored on disk.

`FastFile` mode works best when the data is read sequentially. Augmented manifest files aren't supported. The startup time is lower when there are fewer files in the S3 bucket provided.

Type: String

Valid Values: `Pipe | File | FastFile`

Required: Yes

HyperParameters

The hyperparameters used for the training job.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 100 items.

Key Length Constraints: Minimum length of 0. Maximum length of 256.

Key Pattern: `.*`

Value Length Constraints: Minimum length of 0. Maximum length of 2500.

Value Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingJobStatusCounters

Service: Amazon SageMaker Service

The numbers of training jobs launched by a hyperparameter tuning job, categorized by status.

Contents

Completed

The number of completed training jobs launched by the hyperparameter tuning job.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

InProgress

The number of in-progress training jobs launched by a hyperparameter tuning job.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

NonRetryableError

The number of training jobs that failed and can't be retried. A failed training job can't be retried if it failed because a client error occurred.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

RetryableError

The number of training jobs that failed, but can be retried. A failed training job can be retried only if it failed because an internal service error occurred.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

Stopped

The number of training jobs launched by a hyperparameter tuning job that were manually stopped.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingJobStepMetadata

Service: Amazon SageMaker Service

Metadata for a training job step.

Contents

Arn

The Amazon Resource Name (ARN) of the training job that was run by this step execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingJobSummary

Service: Amazon SageMaker Service

Provides summary information about a training job.

Contents

CreationTime

A timestamp that shows when the training job was created.

Type: Timestamp

Required: Yes

TrainingJobArn

The Amazon Resource Name (ARN) of the training job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-job/[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

TrainingJobName

The name of the training job that you want a summary for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

TrainingJobStatus

The status of the training job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting`

Required: Yes

LastModifiedTime

Timestamp when the training job was last modified.

Type: Timestamp

Required: No

SecondaryStatus

The secondary status of the training job.

Type: String

Valid Values: `Starting | LaunchingMLInstances | PreparingTrainingStack | Downloading | DownloadingTrainingImage | Training | Uploading | Stopping | Stopped | MaxRuntimeExceeded | Completed | Failed | Interrupted | MaxWaitTimeExceeded | Updating | Restarting | Pending`

Required: No

TrainingEndTime

A timestamp that shows when the training job ended. This field is set only if the training job has one of the terminal statuses (`Completed`, `Failed`, or `Stopped`).

Type: Timestamp

Required: No

TrainingPlanArn

The Amazon Resource Name (ARN); of the training plan associated with this training job.

For more information about how to reserve GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Required: No

WarmPoolStatus

The status of the warm pool associated with the training job.

Type: WarmPoolStatus object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingPlanExtension

Service: Amazon SageMaker Service

Details about an extension to a training plan, including the offering ID, dates, status, and cost information.

Contents

TrainingPlanExtensionOfferingId

The unique identifier of the extension offering that was used to create this extension.

Type: String

Required: Yes

AvailabilityZone

The Availability Zone of the extension.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

AvailabilityZoneId

The Availability Zone ID of the extension.

Type: String

Pattern: `[a-z]{3}\d-az\d`

Required: No

CurrencyCode

The currency code for the upfront fee (e.g., USD).

Type: String

Required: No

DurationHours

The duration of the extension in hours.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 4368.

Required: No

EndDate

The end date of the extension period.

Type: Timestamp

Required: No

ExtendedAt

The timestamp when the extension was created.

Type: Timestamp

Required: No

PaymentStatus

The payment processing status of the extension.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

StartDate

The start date of the extension period.

Type: Timestamp

Required: No

Status

The current status of the extension (e.g., Pending, Active, Scheduled, Failed, Expired).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

UpfrontFee

The upfront fee for the extension.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingPlanExtensionOffering

Service: Amazon SageMaker Service

Details about an available extension offering for a training plan. Use the offering ID with the `ExtendTrainingPlan` API to extend a training plan.

Contents

TrainingPlanExtensionOfferingId

The unique identifier for this extension offering.

Type: String

Required: Yes

AvailabilityZone

The Availability Zone for this extension offering.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

CurrencyCode

The currency code for the upfront fee (e.g., USD).

Type: String

Required: No

DurationHours

The duration of this extension offering in hours.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 4368.

Required: No

EndDate

The end date of this extension offering.

Type: Timestamp

Required: No

StartDate

The start date of this extension offering.

Type: Timestamp

Required: No

UpfrontFee

The upfront fee for this extension offering.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingPlanFilter

Service: Amazon SageMaker Service

A filter to apply when listing or searching for training plans.

For more information about how to reserve GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Contents

Name

The name of the filter field (e.g., Status, InstanceType).

Type: String

Valid Values: `Status`

Required: Yes

Value

The value to filter by for the specified field.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingPlanOffering

Service: Amazon SageMaker Service

Details about a training plan offering.

For more information about how to reserve GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Contents

TargetResources

The target resources (e.g., SageMaker Training Jobs, SageMaker HyperPod, SageMaker Endpoints, Studio apps) for this training plan offering.

Training plans are specific to their target resource.

- A training plan designed for SageMaker training jobs can only be used to schedule and run training jobs.
- A training plan for HyperPod clusters can be used exclusively to provide compute resources to a cluster's instance group.
- A training plan for SageMaker endpoints can be used exclusively to provide compute resources to SageMaker endpoints for model deployment.
- A training plan for Studio apps can be used to launch JupyterLab and Code Editor apps on reserved training plan capacity.

Type: Array of strings

Array Members: Minimum number of 1 item.

Valid Values: `training-job | hyperpod-cluster | endpoint | studio-apps`

Required: Yes

TrainingPlanOfferingId

The unique identifier for this training plan offering.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-z0-9\-]+`

Required: Yes

CurrencyCode

The currency code for the upfront fee (e.g., USD).

Type: String

Required: No

DurationHours

The number of whole hours in the total duration for this training plan offering.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 87600.

Required: No

DurationMinutes

The additional minutes beyond whole hours in the total duration for this training plan offering.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 59.

Required: No

RequestedEndTimeBefore

The requested end time that the user specified when searching for the training plan offering.

Type: Timestamp

Required: No

RequestedStartTimeAfter

The requested start time that the user specified when searching for the training plan offering.

Type: Timestamp

Required: No

ReservedCapacityOfferings

A list of reserved capacity offerings associated with this training plan offering.

Type: Array of ReservedCapacityOffering objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

UpfrontFee

The upfront fee for this training plan offering.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingPlanSummary

Service: Amazon SageMaker Service

Details of the training plan.

For more information about how to reserve GPU capacity for your SageMaker HyperPod clusters using Amazon SageMaker Training Plan, see `CreateTrainingPlan` .

Contents

Status

The current status of the training plan (e.g., Pending, Active, Expired). To see the complete list of status values available for a training plan, refer to the `Status` attribute within the `TrainingPlanSummary` object.

Type: String

Valid Values: `Pending | Active | Scheduled | Expired | Failed`

Required: Yes

TrainingPlanArn

The Amazon Resource Name (ARN); of the training plan.

Type: String

Length Constraints: Minimum length of 50. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:training-plan/.*`

Required: Yes

TrainingPlanName

The name of the training plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,63}`

Required: Yes

AvailableInstanceCount

The number of instances currently available for use in this training plan.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

CurrencyCode

The currency code for the upfront fee (e.g., USD).

Type: String

Required: No

DurationHours

The number of whole hours in the total duration for this training plan.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 87600.

Required: No

DurationMinutes

The additional minutes beyond whole hours in the total duration for this training plan.

Type: Long

Valid Range: Minimum value of 0. Maximum value of 59.

Required: No

EndTime

The end time of the training plan.

Type: Timestamp

Required: No

InUseInstanceCount

The number of instances currently in use from this training plan.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

ReservedCapacitySummaries

A list of reserved capacities associated with this training plan, including details such as instance types, counts, and availability zones.

Type: Array of ReservedCapacitySummary objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

StartTime

The start time of the training plan.

Type: Timestamp

Required: No

StatusMessage

A message providing additional information about the current status of the training plan.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

TargetResources

The target resources (e.g., training jobs, HyperPod clusters, Endpoints, Studio apps) that can use this training plan.

Training plans are specific to their target resource.

- A training plan designed for SageMaker training jobs can only be used to schedule and run training jobs.
- A training plan for HyperPod clusters can be used exclusively to provide compute resources to a cluster's instance group.
- A training plan for SageMaker endpoints can be used exclusively to provide compute resources to SageMaker endpoints for model deployment.
- A training plan for Studio apps can be used to launch JupyterLab and Code Editor apps on reserved training plan capacity.

Type: Array of strings

Array Members: Minimum number of 1 item.

Valid Values: `training-job | hyperpod-cluster | endpoint | studio-apps`

Required: No

TotalInstanceCount

The total number of instances reserved in this training plan.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

TotalUltraServerCount

The total number of UltraServers allocated to this training plan.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

UpfrontFee

The upfront fee for the training plan.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingProgressInfo

Service: Amazon SageMaker Service

The serverless training job progress information.

Contents

CurrentEpoch

The current epoch number.

Type: Long

Valid Range: Minimum value of 0.

Required: No

CurrentStep

The current step number.

Type: Long

Valid Range: Minimum value of 0.

Required: No

MaxEpoch

The maximum number of epochs for this job.

Type: Long

Valid Range: Minimum value of 0.

Required: No

TotalStepCountPerEpoch

The total step count per epoch.

Type: Long

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingRepositoryAuthConfig

Service: Amazon SageMaker Service

An object containing authentication information for a private Docker registry.

Contents

TrainingRepositoryCredentialsProviderArn

The Amazon Resource Name (ARN) of an AWS Lambda function used to give SageMaker access credentials to your private Docker registry.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:[\p{Alnum}\-]+:lambda:[\p{Alnum}\-]+:[0-9]{12}:function:.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingSpecification

Service: Amazon SageMaker Service

Defines how the algorithm is used for a training job.

Contents

SupportedTrainingInstanceTypes

A list of the instance types that this algorithm can use for training.

Type: Array of strings

Valid Values: `ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge | ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.p3dn.24xlarge | ml.p4d.24xlarge | ml.p4de.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.c5n.xlarge | ml.c5n.2xlarge | ml.c5n.4xlarge | ml.c5n.9xlarge | ml.c5n.18xlarge | ml.g5.xlarge | ml.g5.2xlarge | ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.16xlarge | ml.g5.12xlarge | ml.g5.24xlarge | ml.g5.48xlarge | ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge | ml.g6.16xlarge | ml.g6.12xlarge | ml.g6.24xlarge | ml.g6.48xlarge | ml.g6e.xlarge | ml.g6e.2xlarge | ml.g6e.4xlarge | ml.g6e.8xlarge | ml.g6e.16xlarge | ml.g6e.12xlarge | ml.g6e.24xlarge | ml.g6e.48xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge | ml.trn1n.32xlarge | ml.trn2.48xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.8xlarge | ml.c6i.4xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.r5d.large | ml.r5d.xlarge | ml.r5d.2xlarge | ml.r5d.4xlarge | ml.r5d.8xlarge | ml.r5d.12xlarge | ml.r5d.16xlarge | ml.r5d.24xlarge | ml.t3.medium | ml.t3.large | ml.t3.xlarge | ml.t3.2xlarge | ml.r5.large`

```
| ml.r5.xlarge | ml.r5.2xlarge | ml.r5.4xlarge | ml.r5.8xlarge |
ml.r5.12xlarge | ml.r5.16xlarge | ml.r5.24xlarge | ml.p6-b200.48xlarge
| ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge | ml.m7i.4xlarge |
ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge | ml.m7i.24xlarge
| ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge | ml.c7i.2xlarge |
ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge | ml.c7i.16xlarge
| ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large | ml.r7i.xlarge
| ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge | ml.r7i.12xlarge
| ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge | ml.p6e-
gb200.36xlarge | ml.p5.4xlarge | ml.p6-b300.48xlarge | ml.g7e.2xlarge
| ml.g7e.4xlarge | ml.g7e.8xlarge | ml.g7e.12xlarge | ml.g7e.24xlarge |
ml.g7e.48xlarge
```

Required: Yes

TrainingChannels

A list of `ChannelSpecification` objects, which specify the input sources to be used by the algorithm.

Type: Array of ChannelSpecification objects

Array Members: Minimum number of 1 item. Maximum number of 8 items.

Required: Yes

TrainingImage

The Amazon ECR registry path of the Docker image that contains the training algorithm.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 255.

Pattern: `[\S]+`

Required: Yes

AdditionalS3DataSource

The additional data source used during the training job.

Type: AdditionalS3DataSource object

Required: No

MetricDefinitions

A list of `MetricDefinition` objects, which are used for parsing metrics generated by the algorithm.

Type: Array of MetricDefinition objects

Array Members: Minimum number of 0 items. Maximum number of 40 items.

Required: No

SupportedHyperParameters

A list of the `HyperParameterSpecification` objects, that define the supported hyperparameters. This is required if the algorithm supports automatic model tuning.>

Type: Array of HyperParameterSpecification objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Required: No

SupportedTuningJobObjectiveMetrics

A list of the metrics that the algorithm emits that can be used as the objective metric in a hyperparameter tuning job.

Type: Array of HyperParameterTuningJobObjective objects

Required: No

SupportsDistributedTraining

Indicates whether the algorithm supports distributed training. If set to false, buyers can't request more than one instance during training.

Type: Boolean

Required: No

TrainingImageDigest

An MD5 hash of the training algorithm that identifies the Docker image used for training.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 72.

Pattern: `[Ss][Hh][Aa]256:[0-9a-fA-F]{64}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TransformDataSource

Service: Amazon SageMaker Service

Describes the location of the channel data.

Contents

S3DataSource

The S3 location of the data source that is associated with a channel.

Type: TransformS3DataSource object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TransformInput

Service: Amazon SageMaker Service

Describes the input source of a transform job and the way the transform job consumes it.

Contents

DataSource

Describes the location of the channel data, which is, the S3 location of the input data that the model can consume.

Type: TransformDataSource object

Required: Yes

CompressionType

If your transform data is compressed, specify the compression type. Amazon SageMaker automatically decompresses the data for the transform job accordingly. The default value is None.

Type: String

Valid Values: `None | Gzip`

Required: No

ContentType

The multipurpose internet mail extension (MIME) type of the data. Amazon SageMaker uses the MIME type with each http call to transfer data to the transform job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

SplitType

The method to use to split the transform job's data files into smaller batches. Splitting is necessary when the total size of each object is too large to fit in a single request. You can also

use data splitting to improve performance by processing multiple concurrent mini-batches. The default value for `SplitType` is `None`, which indicates that input data files are not split, and request payloads contain the entire contents of an input object. Set the value of this parameter to `Line` to split records on a newline character boundary. `SplitType` also supports a number of record-oriented binary data formats. Currently, the supported record formats are:

- RecordIO
- TFRecord

When splitting is enabled, the size of a mini-batch depends on the values of the `BatchStrategy` and `MaxPayloadInMB` parameters. When the value of `BatchStrategy` is `MultiRecord`, Amazon SageMaker sends the maximum number of records in each request, up to the `MaxPayloadInMB` limit. If the value of `BatchStrategy` is `SingleRecord`, Amazon SageMaker sends individual records in each request.

> **Note**
>
> Some data formats represent a record as a binary payload wrapped with extra padding bytes. When splitting is applied to a binary data format, padding is removed if the value of `BatchStrategy` is set to `SingleRecord`. Padding is not removed if the value of `BatchStrategy` is set to `MultiRecord`.
> For more information about `RecordIO`, see Create a Dataset Using RecordIO in the MXNet documentation. For more information about `TFRecord`, see Consuming TFRecord data in the TensorFlow documentation.

Type: String

Valid Values: `None | Line | RecordIO | TFRecord`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

TransformJob

Service: Amazon SageMaker Service

A batch transform job. For information about SageMaker batch transform, see Use Batch Transform.

Contents

AutoMLJobArn

The Amazon Resource Name (ARN) of the AutoML job that created the transform job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:automl-job/.*`

Required: No

BatchStrategy

Specifies the number of records to include in a mini-batch for an HTTP inference request. A record is a single unit of input data that inference can be made on. For example, a single line in a CSV file is a record.

Type: String

Valid Values: `MultiRecord | SingleRecord`

Required: No

CreationTime

A timestamp that shows when the transform Job was created.

Type: Timestamp

Required: No

DataCaptureConfig

Configuration to control how SageMaker captures inference data for batch transform jobs.

Type: BatchDataCaptureConfig object

Required: No

DataProcessing

The data structure used to specify the data to be used for inference in a batch transform job and to associate the data that is relevant to the prediction results in the output. The input filter provided allows you to exclude input data that is not needed for inference in a batch transform job. The output filter provided allows you to include input data relevant to interpreting the predictions in the output from the job. For more information, see Associate Prediction Results with their Corresponding Input Records.

Type: DataProcessing object

Required: No

Environment

The environment variables to set in the Docker container. We support up to 16 key and values entries in the map.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 16 items.

Key Length Constraints: Minimum length of 0. Maximum length of 1024.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]{0,1023}`

Value Length Constraints: Minimum length of 0. Maximum length of 10240.

Value Pattern: `[\S\s]*`

Required: No

ExperimentConfig

Associates a SageMaker job as a trial component with an experiment and trial. Specified when you call the following APIs:

- CreateProcessingJob
- CreateTrainingJob
- CreateTransformJob

Type: ExperimentConfig object

Required: No

FailureReason

If the transform job failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

LabelingJobArn

The Amazon Resource Name (ARN) of the labeling job that created the transform job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:labeling-job/.*`

Required: No

MaxConcurrentTransforms

The maximum number of parallel requests that can be sent to each instance in a transform job. If `MaxConcurrentTransforms` is set to 0 or left unset, SageMaker checks the optional execution-parameters to determine the settings for your chosen algorithm. If the execution-parameters endpoint is not enabled, the default value is 1. For built-in algorithms, you don't need to set a value for `MaxConcurrentTransforms`.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

MaxPayloadInMB

The maximum allowed size of the payload, in MB. A payload is the data portion of a record (without metadata). The value in `MaxPayloadInMB` must be greater than, or equal to, the size of a single record. To estimate the size of a record in MB, divide the size of your dataset by the number of records. To ensure that the records fit within the maximum payload size, we recommend using a slightly larger value. The default value is 6 MB. For cases where the payload

might be arbitrarily large and is transmitted using HTTP chunked encoding, set the value to 0. This feature works only in supported algorithms. Currently, SageMaker built-in algorithms do not support HTTP chunked encoding.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

ModelClientConfig

Configures the timeout and maximum number of retries for processing a transform job invocation.

Type: ModelClientConfig object

Required: No

ModelName

The name of the model associated with the transform job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([\-a-zA-Z0-9]*[a-zA-Z0-9])?`

Required: No

Tags

A list of tags associated with the transform job.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TransformEndTime

Indicates when the transform job has been completed, or has stopped or failed. You are billed for the time interval between this time and the value of `TransformStartTime`.

Type: Timestamp

Required: No

TransformInput

Describes the input source of a transform job and the way the transform job consumes it.

Type: TransformInput object

Required: No

TransformJobArn

The Amazon Resource Name (ARN) of the transform job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:transform-job/.*`

Required: No

TransformJobName

The name of the transform job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

TransformJobStatus

The status of the transform job.

Transform job statuses are:

- `InProgress` - The job is in progress.
- `Completed` - The job has completed.
- `Failed` - The transform job has failed. To see the reason for the failure, see the `FailureReason` field in the response to a `DescribeTransformJob` call.

- `Stopping` - The transform job is stopping.
- `Stopped` - The transform job has stopped.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: No

TransformOutput

Describes the results of a transform job.

Type: TransformOutput object

Required: No

TransformResources

Describes the resources, including ML instance types and ML instance count, to use for transform job.

Type: TransformResources object

Required: No

TransformStartTime

Indicates when the transform job starts on ML instances. You are billed for the time interval between this time and the value of `TransformEndTime`.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TransformJobDefinition

Service: Amazon SageMaker Service

Defines the input needed to run a transform job using the inference specification specified in the algorithm.

Contents

TransformInput

A description of the input source and the way the transform job consumes it.

Type: TransformInput object

Required: Yes

TransformOutput

Identifies the Amazon S3 location where you want Amazon SageMaker to save the results from the transform job.

Type: TransformOutput object

Required: Yes

TransformResources

Identifies the ML compute instances for the transform job.

Type: TransformResources object

Required: Yes

BatchStrategy

A string that determines the number of records included in a single mini-batch.

`SingleRecord` means only one record is used per mini-batch. `MultiRecord` means a mini-batch is set to contain as many records that can fit within the `MaxPayloadInMB` limit.

Type: String

Valid Values: `MultiRecord | SingleRecord`

Required: No

Environment

The environment variables to set in the Docker container. We support up to 16 key and values entries in the map.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 16 items.

Key Length Constraints: Minimum length of 0. Maximum length of 1024.

Key Pattern: `[a-zA-Z_][a-zA-Z0-9_]{0,1023}`

Value Length Constraints: Minimum length of 0. Maximum length of 10240.

Value Pattern: `[\S\s]*`

Required: No

MaxConcurrentTransforms

The maximum number of parallel requests that can be sent to each instance in a transform job. The default value is 1.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

MaxPayloadInMB

The maximum payload size allowed, in MB. A payload is the data portion of a record (without metadata).

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TransformJobStepMetadata

Service: Amazon SageMaker Service

Metadata for a transform job step.

Contents

Arn

The Amazon Resource Name (ARN) of the transform job that was run by this step execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:transform-job/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TransformJobSummary

Service: Amazon SageMaker Service

Provides a summary of a transform job. Multiple `TransformJobSummary` objects are returned as a list after in response to a ListTransformJobs call.

Contents

CreationTime

A timestamp that shows when the transform Job was created.

Type: Timestamp

Required: Yes

TransformJobArn

The Amazon Resource Name (ARN) of the transform job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:transform-job/.*`

Required: Yes

TransformJobName

The name of the transform job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

TransformJobStatus

The status of the transform job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: Yes

FailureReason

If the transform job failed, the reason it failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

LastModifiedTime

Indicates when the transform job was last modified.

Type: Timestamp

Required: No

TransformEndTime

Indicates when the transform job ends on compute instances. For successful jobs and stopped jobs, this is the exact time recorded after the results are uploaded. For failed jobs, this is when Amazon SageMaker detected that the job failed.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TransformOutput

Service: Amazon SageMaker Service

Describes the results of a transform job.

Contents

S3OutputPath

The Amazon S3 path where you want Amazon SageMaker to store the results of the transform job. For example, `s3://bucket-name/key-name-prefix`.

For every S3 object used as input for the transform job, batch transform stores the transformed data with an `.out` suffix in a corresponding subfolder in the location in the output prefix. For example, for the input data stored at `s3://bucket-name/input-name-prefix/dataset01/data.csv`, batch transform stores the transformed data at `s3://bucket-name/output-name-prefix/input-name-prefix/data.csv.out`. Batch transform doesn't upload partially processed objects. For an input S3 object that contains multiple records, it creates an `.out` file only if the transform job succeeds on the entire file. When the input contains multiple S3 objects, the batch transform job processes the listed S3 objects and uploads only the output for successfully processed objects. If any object fails in the transform job batch transform marks the job as failed to prompt investigation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

Accept

The MIME type used to specify the output data. Amazon SageMaker uses the MIME type with each http call to transfer data from the transform job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `.*`

Required: No

AssembleWith

Defines how to assemble the results of the transform job as a single S3 object. Choose a format that is most convenient to you. To concatenate the results in binary format, specify `None`. To add a newline character at the end of every transformed record, specify `Line`.

Type: String

Valid Values: `None` | `Line`

Required: No

KmsKeyId

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt the model artifacts at rest using Amazon S3 server-side encryption. The `KmsKeyId` can be any of the following formats:

- Key ID: `1234abcd-12ab-34cd-56ef-1234567890ab`
- Key ARN: `arn:aws:kms:us-west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab`
- Alias name: `alias/ExampleAlias`
- Alias name ARN: `arn:aws:kms:us-west-2:111122223333:alias/ExampleAlias`

If you don't provide a KMS key ID, Amazon SageMaker uses the default KMS key for Amazon S3 for your role's account. For more information, see KMS-Managed Encryption Keys in the *Amazon Simple Storage Service Developer Guide.*

The KMS key policy must grant permission to the IAM role that you specify in your CreateModel request. For more information, see Using Key Policies in AWS KMS in the *AWS Key Management Service Developer Guide*.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TransformResources

Service: Amazon SageMaker Service

Describes the resources, including ML instance types and ML instance count, to use for transform job.

Contents

InstanceCount

The number of ML compute instances to use in the transform job. The default value is 1, and the maximum is 100. For distributed transform jobs, specify a value greater than 1.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

InstanceType

The ML compute instance type for the transform job. If you are using built-in algorithms to transform moderately sized datasets, we recommend using ml.m4.xlarge or ml.m5.largeinstance types.

Type: String

Valid Values: ml.m4.xlarge | ml.m4.2xlarge | ml.m4.4xlarge | ml.m4.10xlarge | ml.m4.16xlarge | ml.c4.xlarge | ml.c4.2xlarge | ml.c4.4xlarge | ml.c4.8xlarge | ml.p2.xlarge | ml.p2.8xlarge | ml.p2.16xlarge | ml.p3.2xlarge | ml.p3.8xlarge | ml.p3.16xlarge | ml.c5.xlarge | ml.c5.2xlarge | ml.c5.4xlarge | ml.c5.9xlarge | ml.c5.18xlarge | ml.m5.large | ml.m5.xlarge | ml.m5.2xlarge | ml.m5.4xlarge | ml.m5.12xlarge | ml.m5.24xlarge | ml.m6i.large | ml.m6i.xlarge | ml.m6i.2xlarge | ml.m6i.4xlarge | ml.m6i.8xlarge | ml.m6i.12xlarge | ml.m6i.16xlarge | ml.m6i.24xlarge | ml.m6i.32xlarge | ml.c6i.large | ml.c6i.xlarge | ml.c6i.2xlarge | ml.c6i.4xlarge | ml.c6i.8xlarge | ml.c6i.12xlarge | ml.c6i.16xlarge | ml.c6i.24xlarge | ml.c6i.32xlarge | ml.r6i.large | ml.r6i.xlarge | ml.r6i.2xlarge | ml.r6i.4xlarge | ml.r6i.8xlarge | ml.r6i.12xlarge | ml.r6i.16xlarge | ml.r6i.24xlarge | ml.r6i.32xlarge | ml.m7i.large | ml.m7i.xlarge | ml.m7i.2xlarge |

```
ml.m7i.4xlarge | ml.m7i.8xlarge | ml.m7i.12xlarge | ml.m7i.16xlarge
| ml.m7i.24xlarge | ml.m7i.48xlarge | ml.c7i.large | ml.c7i.xlarge
| ml.c7i.2xlarge | ml.c7i.4xlarge | ml.c7i.8xlarge | ml.c7i.12xlarge
| ml.c7i.16xlarge | ml.c7i.24xlarge | ml.c7i.48xlarge | ml.r7i.large
| ml.r7i.xlarge | ml.r7i.2xlarge | ml.r7i.4xlarge | ml.r7i.8xlarge |
ml.r7i.12xlarge | ml.r7i.16xlarge | ml.r7i.24xlarge | ml.r7i.48xlarge
| ml.g4dn.xlarge | ml.g4dn.2xlarge | ml.g4dn.4xlarge | ml.g4dn.8xlarge
| ml.g4dn.12xlarge | ml.g4dn.16xlarge | ml.g5.xlarge | ml.g5.2xlarge
| ml.g5.4xlarge | ml.g5.8xlarge | ml.g5.12xlarge | ml.g5.16xlarge |
ml.g5.24xlarge | ml.g5.48xlarge | ml.trn1.2xlarge | ml.trn1.32xlarge |
ml.inf2.xlarge | ml.inf2.8xlarge | ml.inf2.24xlarge | ml.inf2.48xlarge
| ml.g6.xlarge | ml.g6.2xlarge | ml.g6.4xlarge | ml.g6.8xlarge |
ml.g6.12xlarge | ml.g6.16xlarge | ml.g6.24xlarge | ml.g6.48xlarge
```

Required: Yes

TransformAmiVersion

Specifies an option from a collection of preconfigured Amazon Machine Image (AMI) images. Each image is configured by AWS with a set of software and driver versions.

al2-ami-sagemaker-batch-gpu-470

- Accelerator: GPU
- NVIDIA driver version: 470

al2-ami-sagemaker-batch-gpu-535

- Accelerator: GPU
- NVIDIA driver version: 535

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]+(-[a-zA-Z0-9]+)*`

Required: No

VolumeKmsKeyId

The AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to encrypt model data on the storage volume attached to the ML compute instance(s) that run the batch transform job.

> **Note**
>
> Certain Nitro-based instances include local storage, dependent on the instance type. Local storage volumes are encrypted using a hardware module on the instance. You can't request a `VolumeKmsKeyId` when using an instance type with local storage.
> For a list of instance types that support local instance storage, see Instance Store Volumes.
> For more information about local instance storage encryption, see SSD Instance Store Volumes.

The `VolumeKmsKeyId` can be any of the following formats:

- Key ID: `1234abcd-12ab-34cd-56ef-1234567890ab`
- Key ARN: `arn:aws:kms:us-west-2:111122223333:key/1234abcd-12ab-34cd-56ef-1234567890ab`
- Alias name: `alias/ExampleAlias`
- Alias name ARN: `arn:aws:kms:us-west-2:111122223333:alias/ExampleAlias`

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TransformS3DataSource

Service: Amazon SageMaker Service

Describes the S3 data source.

Contents

S3DataType

If you choose `S3Prefix`, `S3Uri` identifies a key name prefix. Amazon SageMaker uses all objects with the specified key name prefix for batch transform.

If you choose `ManifestFile`, `S3Uri` identifies an object that is a manifest file containing a list of object keys that you want Amazon SageMaker to use for batch transform.

The following values are compatible: `ManifestFile`, `S3Prefix`

The following value is not compatible: `AugmentedManifestFile`

Type: String

Valid Values: `ManifestFile | S3Prefix | AugmentedManifestFile | Converse`

Required: Yes

S3Uri

Depending on the value specified for the `S3DataType`, identifies either a key name prefix or a manifest. For example:

- A key name prefix might look like this: `s3://bucketname/exampleprefix/`.
- A manifest might look like this: `s3://bucketname/example.manifest`

  The manifest is an S3 object which is a JSON file with the following format:

  ```
  [ {"prefix": "s3://customer_bucket/some/prefix/"},

  "relative/path/to/custdata-1",

  "relative/path/custdata-2",

  ...

  "relative/path/custdata-N"
  ```

```
]
```

The preceding JSON matches the following `S3Uris`:

```
s3://customer_bucket/some/prefix/relative/path/to/custdata-1

s3://customer_bucket/some/prefix/relative/path/custdata-2

...

s3://customer_bucket/some/prefix/relative/path/custdata-N
```

The complete set of `S3Uris` in this manifest constitutes the input data for the channel for this datasource. The object that each `S3Uris` points to must be readable by the IAM role that Amazon SageMaker uses to perform tasks on your behalf.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Trial

Service: Amazon SageMaker Service

The properties of a trial as returned by the Search API.

Contents

CreatedBy

Who created the trial.

Type: UserContext object

Required: No

CreationTime

When the trial was created.

Type: Timestamp

Required: No

DisplayName

The name of the trial as displayed. If `DisplayName` isn't specified, `TrialName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

ExperimentName

The name of the experiment the trial is part of.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

LastModifiedTime

Who last modified the trial.

Type: Timestamp

Required: No

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Required: No

Source

The source of the trial.

Type: TrialSource object

Required: No

Tags

The list of tags that are associated with the trial. You can use Search API to search on the tags.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TrialArn

The Amazon Resource Name (ARN) of the trial.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial/.*`

Required: No

TrialComponentSummaries

A list of the components associated with the trial. For each component, a summary of the component's properties is included.

Type: Array of TrialComponentSimpleSummary objects

Required: No

TrialName

The name of the trial.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrialComponent

Service: Amazon SageMaker Service

The properties of a trial component as returned by the Search API.

Contents

CreatedBy

Who created the trial component.

Type: UserContext object

Required: No

CreationTime

When the component was created.

Type: Timestamp

Required: No

DisplayName

The name of the component as displayed. If `DisplayName` isn't specified, `TrialComponentName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

EndTime

When the component ended.

Type: Timestamp

Required: No

InputArtifacts

The input artifacts of the component.

Type: String to TrialComponentArtifact object map

Map Entries: Minimum number of 0 items. Maximum number of 60 items.

Key Length Constraints: Minimum length of 0. Maximum length of 128.

Key Pattern: `.*`

Required: No

LastModifiedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

LastModifiedTime

When the component was last modified.

Type: Timestamp

Required: No

LineageGroupArn

The Amazon Resource Name (ARN) of the lineage group resource.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:lineage-group/.*`

Required: No

MetadataProperties

Metadata properties of the tracking entity, trial, or trial component.

Type: MetadataProperties object

Required: No

Metrics

The metrics for the component.

Type: Array of TrialComponentMetricSummary objects

Required: No

OutputArtifacts

The output artifacts of the component.

Type: String to TrialComponentArtifact object map

Map Entries: Minimum number of 0 items. Maximum number of 60 items.

Key Length Constraints: Minimum length of 0. Maximum length of 128.

Key Pattern: `.*`

Required: No

Parameters

The hyperparameters of the component.

Type: String to TrialComponentParameterValue object map

Map Entries: Minimum number of 0 items. Maximum number of 300 items.

Key Length Constraints: Minimum length of 0. Maximum length of 320.

Key Pattern: `.*`

Required: No

Parents

An array of the parents of the component. A parent is a trial the component is associated with and the experiment the trial is part of. A component might not have any parents.

Type: Array of Parent objects

Required: No

RunName

The name of the experiment run.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

Source

The Amazon Resource Name (ARN) and job type of the source of the component.

Type: TrialComponentSource object

Required: No

SourceDetail

Details of the source of the component.

Type: TrialComponentSourceDetail object

Required: No

StartTime

When the component started.

Type: Timestamp

Required: No

Status

The status of the trial component.

Type: TrialComponentStatus object

Required: No

Tags

The list of tags that are associated with the component. You can use Search API to search on the tags.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

TrialComponentArn

The Amazon Resource Name (ARN) of the trial component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial-component/.*`

Required: No

TrialComponentName

The name of the trial component.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrialComponentArtifact

Service: Amazon SageMaker Service

Represents an input or output artifact of a trial component. You specify `TrialComponentArtifact` as part of the `InputArtifacts` and `OutputArtifacts` parameters in the CreateTrialComponent request.

Examples of input artifacts are datasets, algorithms, hyperparameters, source code, and instance types. Examples of output artifacts are metrics, snapshots, logs, and images.

Contents

Value

The location of the artifact.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

MediaType

The media type of the artifact, which indicates the type of data in the artifact file. The media type consists of a *type* and a *subtype* concatenated with a slash (/) character, for example, text/csv, image/jpeg, and s3/uri. The type specifies the category of the media. The subtype specifies the kind of data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[-\w]+\/[-\w+]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrialComponentMetricSummary

Service: Amazon SageMaker Service

A summary of the metrics of a trial component.

Contents

Avg

The average value of the metric.

Type: Double

Required: No

Count

The number of samples used to generate the metric.

Type: Integer

Required: No

Last

The most recent value of the metric.

Type: Double

Required: No

Max

The maximum value of the metric.

Type: Double

Required: No

MetricName

The name of the metric.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: No

Min

The minimum value of the metric.

Type: Double

Required: No

SourceArn

The Amazon Resource Name (ARN) of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:.*`

Required: No

StdDev

The standard deviation of the metric.

Type: Double

Required: No

TimeStamp

When the metric was last updated.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrialComponentParameterValue

Service: Amazon SageMaker Service

The value of a hyperparameter. Only one of `NumberValue` or `StringValue` can be specified.

This object is specified in the CreateTrialComponent request.

Contents

NumberValue

The numeric value of a numeric hyperparameter. If you specify a value for this parameter, you can't specify the `StringValue` parameter.

Type: Double

Required: No

StringValue

The string value of a categorical hyperparameter. If you specify a value for this parameter, you can't specify the `NumberValue` parameter.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2500.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrialComponentSimpleSummary

Service: Amazon SageMaker Service

A short summary of a trial component.

Contents

CreatedBy

Information about the user who created or modified a SageMaker resource.

Type: UserContext object

Required: No

CreationTime

When the component was created.

Type: Timestamp

Required: No

TrialComponentArn

The Amazon Resource Name (ARN) of the trial component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial-component/.*`

Required: No

TrialComponentName

The name of the trial component.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

TrialComponentSource

The Amazon Resource Name (ARN) and job type of the source of a trial component.

Type: TrialComponentSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrialComponentSource

Service: Amazon SageMaker Service

The Amazon Resource Name (ARN) and job type of the source of a trial component.

Contents

SourceArn

The source Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:.*`

Required: Yes

SourceType

The source job type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrialComponentSourceDetail

Service: Amazon SageMaker Service

Detailed information about the source of a trial component. Either `ProcessingJob` or `TrainingJob` is returned.

Contents

ProcessingJob

Information about a processing job that's the source of a trial component.

Type: ProcessingJob object

Required: No

SourceArn

The Amazon Resource Name (ARN) of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:.*`

Required: No

TrainingJob

Information about a training job that's the source of a trial component.

Type: TrainingJob object

Required: No

TransformJob

Information about a transform job that's the source of a trial component.

Type: TransformJob object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrialComponentStatus

Service: Amazon SageMaker Service

The status of the trial component.

Contents

Message

If the component failed, a message describing why.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `.*`

Required: No

PrimaryStatus

The status of the trial component.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrialComponentSummary

Service: Amazon SageMaker Service

A summary of the properties of a trial component. To get all the properties, call the DescribeTrialComponent API and provide the `TrialComponentName`.

Contents

CreatedBy

Who created the trial component.

Type: UserContext object

Required: No

CreationTime

When the component was created.

Type: Timestamp

Required: No

DisplayName

The name of the component as displayed. If `DisplayName` isn't specified, `TrialComponentName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

EndTime

When the component ended.

Type: Timestamp

Required: No

LastModifiedBy

Who last modified the component.

Type: UserContext object

Required: No

LastModifiedTime

When the component was last modified.

Type: Timestamp

Required: No

StartTime

When the component started.

Type: Timestamp

Required: No

Status

The status of the component. States include:

- InProgress
- Completed
- Failed

Type: TrialComponentStatus object

Required: No

TrialComponentArn

The Amazon Resource Name (ARN) of the trial component.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial-component/.*`

Required: No

TrialComponentName

The name of the trial component.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

TrialComponentSource

The Amazon Resource Name (ARN) and job type of the source of a trial component.

Type: TrialComponentSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrialSource

Service: Amazon SageMaker Service

The source of the trial.

Contents

SourceArn

The Amazon Resource Name (ARN) of the source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:.*`

Required: Yes

SourceType

The source job type.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrialSummary

Service: Amazon SageMaker Service

A summary of the properties of a trial. To get the complete set of properties, call the DescribeTrial API and provide the `TrialName`.

Contents

CreationTime

When the trial was created.

Type: Timestamp

Required: No

DisplayName

The name of the trial as displayed. If `DisplayName` isn't specified, `TrialName` is displayed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

LastModifiedTime

When the trial was last modified.

Type: Timestamp

Required: No

TrialArn

The Amazon Resource Name (ARN) of the trial.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:experiment-trial/.*`

Required: No

TrialName

The name of the trial.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 120.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,119}`

Required: No

TrialSource

The source of the trial.

Type: TrialSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrustedIdentityPropagationSettings

Service: Amazon SageMaker Service

The Trusted Identity Propagation (TIP) settings for the SageMaker domain. These settings determine how user identities from IAM Identity Center are propagated through the domain to TIP enabled AWS services.

Contents

Status

The status of Trusted Identity Propagation (TIP) at the SageMaker domain level.

When disabled, standard IAM role-based access is used.

When enabled:

- User identities from IAM Identity Center are propagated through the application to TIP enabled AWS services.
- New applications or existing applications that are automatically patched, will use the domain level configuration.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TtlDuration

Service: Amazon SageMaker Service

Time to live duration, where the record is hard deleted after the expiration time is reached; `ExpiresAt` = `EventTime` + `TtlDuration`. For information on HardDelete, see the DeleteRecord API in the Amazon SageMaker API Reference guide.

Contents

Unit

`TtlDuration` time unit.

Type: String

Valid Values: `Seconds | Minutes | Hours | Days | Weeks`

Required: No

Value

`TtlDuration` time value.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TuningJobCompletionCriteria

Service: Amazon SageMaker Service

The job completion criteria.

Contents

BestObjectiveNotImproving

A flag to stop your hyperparameter tuning job if model performance fails to improve as evaluated against an objective function.

Type: BestObjectiveNotImproving object

Required: No

ConvergenceDetected

A flag to top your hyperparameter tuning job if automatic model tuning (AMT) has detected that your model has converged as evaluated against your objective function.

Type: ConvergenceDetected object

Required: No

TargetObjectiveMetricValue

The value of the objective metric.

Type: Float

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TuningJobStepMetaData

Service: Amazon SageMaker Service

Metadata for a tuning step.

Contents

Arn

The Amazon Resource Name (ARN) of the tuning job that was run by this step execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:hyper-parameter-tuning-job/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UiConfig

Service: Amazon SageMaker Service

Provided configuration information for the worker UI for a labeling job. Provide either `HumanTaskUiArn` or `UiTemplateS3Uri`.

For named entity recognition, 3D point cloud and video frame labeling jobs, use `HumanTaskUiArn`.

For all other Ground Truth built-in task types and custom task types, use `UiTemplateS3Uri` to specify the location of a worker task template in Amazon S3.

Contents

HumanTaskUiArn

The ARN of the worker task template used to render the worker UI and tools for labeling job tasks.

Use this parameter when you are creating a labeling job for named entity recognition, 3D point cloud and video frame labeling jobs. Use your labeling job task type to select one of the following ARNs and use it with this parameter when you create a labeling job. Replace `aws-region` with the AWS Region you are creating your labeling job in. For example, replace `aws-region` with `us-west-1` if you create a labeling job in US West (N. California).

Named Entity Recognition

Use the following `HumanTaskUiArn` for named entity recognition labeling jobs:

`arn:aws:sagemaker:aws-region:394669845002:human-task-ui/NamedEntityRecognition`

3D Point Cloud HumanTaskUiArns

Use this `HumanTaskUiArn` for 3D point cloud object detection and 3D point cloud object detection adjustment labeling jobs.

- `arn:aws:sagemaker:aws-region:394669845002:human-task-ui/PointCloudObjectDetection`

Use this `HumanTaskUiArn` for 3D point cloud object tracking and 3D point cloud object tracking adjustment labeling jobs.

- `arn:aws:sagemaker:aws-region:394669845002:human-task-ui/PointCloudObjectTracking`

Use this `HumanTaskUiArn` for 3D point cloud semantic segmentation and 3D point cloud semantic segmentation adjustment labeling jobs.

- `arn:aws:sagemaker:aws-region:394669845002:human-task-ui/PointCloudSemanticSegmentation`

Video Frame HumanTaskUiArns

Use this `HumanTaskUiArn` for video frame object detection and video frame object detection adjustment labeling jobs.

- `arn:aws:sagemaker:region:394669845002:human-task-ui/VideoObjectDetection`

Use this `HumanTaskUiArn` for video frame object tracking and video frame object tracking adjustment labeling jobs.

- `arn:aws:sagemaker:aws-region:394669845002:human-task-ui/VideoObjectTracking`

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]+:[0-9]{12}:human-task-ui/.*`

Required: No

UiTemplateS3Uri

The Amazon S3 bucket location of the UI template, or worker task template. This is the template used to render the worker UI and tools for labeling job tasks. For more information about the contents of a UI template, see Creating Your Custom Labeling Task Template.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UiTemplate

Service: Amazon SageMaker Service

The Liquid template for the worker user interface.

Contents

Content

The content of the Liquid template for the worker user interface.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128000.

Pattern: `[\S\s]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UiTemplateInfo

Service: Amazon SageMaker Service

Container for user interface template information.

Contents

ContentSha256

The SHA-256 digest of the contents of the template.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128000.

Required: No

Url

The URL for the user interface template.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UltraServer

Service: Amazon SageMaker Service

Represents a high-performance compute server used for distributed training in SageMaker AI. An UltraServer consists of multiple instances within a shared NVLink interconnect domain.

Contents

AvailabilityZone

The name of the Availability Zone where the UltraServer is provisioned.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32.

Pattern: `[a-z]+\-[0-9a-z\-]+`

Required: Yes

InstanceType

The Amazon EC2 instance type used in the UltraServer.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.trn1.32xlarge | ml.trn2.48xlarge | ml.p6-b200.48xlarge | ml.p4de.24xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge | ml.p6-b300.48xlarge`

Required: Yes

TotalInstanceCount

The total number of instances in this UltraServer.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

UltraServerId

The unique identifier for the UltraServer.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `(?!\s*$).+`

Required: Yes

UltraServerType

The type of UltraServer, such as ml.u-p6e-gb200x72.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `ml.[a-z0-9\-.]+`

Required: Yes

AvailableInstanceCount

The number of instances currently available for use in this UltraServer.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

AvailableSpareInstanceCount

The number of available spare instances in the UltraServer.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

ConfiguredSpareInstanceCount

The number of spare instances configured for this UltraServer to provide enhanced resiliency.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

HealthStatus

The overall health status of the UltraServer.

Type: String

Valid Values: `OK | Impaired | Insufficient-Data`

Required: No

InUseInstanceCount

The number of instances currently in use in this UltraServer.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

UnhealthyInstanceCount

The number of instances in this UltraServer that are currently in an unhealthy state.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UltraServerInfo

Service: Amazon SageMaker Service

Contains information about the UltraServer object.

Contents

Id

The unique identifier of the UltraServer.

Type: String

Required: No

Type

The type of the UltraServer.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UltraServerSummary

Service: Amazon SageMaker Service

A summary of UltraServer resources and their current status.

Contents

InstanceType

The Amazon EC2 instance type used in the UltraServer.

Type: String

Valid Values: `ml.p4d.24xlarge | ml.p5.48xlarge | ml.p5e.48xlarge | ml.p5en.48xlarge | ml.trn1.32xlarge | ml.trn2.48xlarge | ml.p6-b200.48xlarge | ml.p4de.24xlarge | ml.p6e-gb200.36xlarge | ml.p5.4xlarge | ml.p6-b300.48xlarge`

Required: Yes

UltraServerType

The type of UltraServer, such as ml.u-p6e-gb200x72.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `ml.[a-z0-9\-.]+`

Required: Yes

AvailableSpareInstanceCount

The number of available spare instances in the UltraServers.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

UltraServerCount

The number of UltraServers of this type.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

UnhealthyInstanceCount

The total number of instances across all UltraServers of this type that are currently in an unhealthy state.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnifiedStudioSettings

Service: Amazon SageMaker Service

The settings that apply to an Amazon SageMaker AI domain when you use it in Amazon SageMaker Unified Studio.

Contents

DomainAccountId

The ID of the AWS account that has the Amazon SageMaker Unified Studio domain. The default value, if you don't specify an ID, is the ID of the account that has the Amazon SageMaker AI domain.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `\d+`

Required: No

DomainId

The ID of the Amazon SageMaker Unified Studio domain associated with this domain.

Type: String

Pattern: `dzd[-_][a-zA-Z0-9_-]{1,36}`

Required: No

DomainRegion

The AWS Region where the domain is located in Amazon SageMaker Unified Studio. The default value, if you don't specify a Region, is the Region where the Amazon SageMaker AI domain is located.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 24.

Required: No

EnvironmentId

The ID of the environment that Amazon SageMaker Unified Studio associates with the domain.

Type: String

Pattern: `[a-zA-Z0-9_-]{1,36}`

Required: No

ProjectId

The ID of the Amazon SageMaker Unified Studio project that corresponds to the domain.

Type: String

Pattern: `[a-zA-Z0-9_-]{1,36}`

Required: No

ProjectS3Path

The location where Amazon S3 stores temporary execution data and other artifacts for the project that corresponds to the domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

SingleSignOnApplicationArn

The ARN of the Amazon DataZone application managed by Amazon SageMaker Unified Studio in the AWS IAM Identity Center.

Type: String

Pattern: `arn:(aws|aws-us-gov|aws-cn|aws-iso|aws-iso-b):sso::[0-9]+:application\/[a-zA-Z0-9-_.]+\/apl-[a-zA-Z0-9]+`

Required: No

StudioWebPortalAccess

Sets whether you can access the domain in Amazon SageMaker Studio:

ENABLED

You can access the domain in Amazon SageMaker Studio. If you migrate the domain to Amazon SageMaker Unified Studio, you can access it in both studio interfaces.

DISABLED

You can't access the domain in Amazon SageMaker Studio. If you migrate the domain to Amazon SageMaker Unified Studio, you can access it only in that studio interface.

To migrate a domain to Amazon SageMaker Unified Studio, you specify the UnifiedStudioSettings data type when you use the UpdateDomain action.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UpdateClusterSoftwareInstanceGroupSpecification

Service: Amazon SageMaker Service

The configuration that describes specifications of the instance groups to update.

Contents

InstanceGroupName

The name of the instance group to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UpdateTemplateProvider

Service: Amazon SageMaker Service

Contains configuration details for updating an existing template provider in the project.

Contents

CfnTemplateProvider

The CloudFormation template provider configuration to update.

Type: CfnUpdateTemplateProvider object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

USD

Service: Amazon SageMaker Service

Represents an amount of money in United States dollars.

Contents

Cents

The fractional portion, in cents, of the amount.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 99.

Required: No

Dollars

The whole number of dollars in the amount.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 2.

Required: No

TenthFractionsOfACent

Fractions of a cent, in tenths.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 9.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

UserContext

Service: Amazon SageMaker Service

Information about the user who created or modified a SageMaker resource.

Contents

DomainId

The domain associated with the user.

Type: String

Required: No

IamIdentity

The IAM Identity details associated with the user. These details are associated with model package groups, model packages, and project entities only.

Type: IamIdentity object

Required: No

UserProfileArn

The Amazon Resource Name (ARN) of the user's profile.

Type: String

Required: No

UserProfileName

The name of the user's profile.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UserProfileDetails

Service: Amazon SageMaker Service

The user profile details.

Contents

CreationTime

The creation time.

Type: Timestamp

Required: No

DomainId

The domain ID.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `d-(-*[a-z0-9]){1,61}`

Required: No

LastModifiedTime

The last modified time.

Type: Timestamp

Required: No

Status

The status.

Type: String

Valid Values: `Deleting | Failed | InService | Pending | Updating | Update_Failed | Delete_Failed`

Required: No

UserProfileName

The user profile name.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UserSettings

Service: Amazon SageMaker Service

A collection of settings that apply to users in a domain. These settings are specified when the `CreateUserProfile` API is called, and as `DefaultUserSettings` when the `CreateDomain` API is called.

`SecurityGroups` is aggregated when specified in both calls. For all other settings in `UserSettings`, the values specified in `CreateUserProfile` take precedence over those specified in `CreateDomain`.

Contents

AutoMountHomeEFS

Indicates whether auto-mounting of an EFS volume is supported for the user profile. The `DefaultAsDomain` value is only supported for user profiles. Do not use the `DefaultAsDomain` value when setting this parameter for a domain.

SageMaker applies this setting only to private spaces that the user creates in the domain. SageMaker doesn't apply this setting to shared spaces.

Type: String

Valid Values: `Enabled | Disabled | DefaultAsDomain`

Required: No

CanvasAppSettings

The Canvas app settings.

SageMaker applies these settings only to private spaces that SageMaker creates for the Canvas app.

Type: CanvasAppSettings object

Required: No

CodeEditorAppSettings

The Code Editor application settings.

SageMaker applies these settings only to private spaces that the user creates in the domain. SageMaker doesn't apply these settings to shared spaces.

Type: CodeEditorAppSettings object

Required: No

CustomFileSystemConfigs

The settings for assigning a custom file system to a user profile. Permitted users can access this file system in Amazon SageMaker AI Studio.

SageMaker applies these settings only to private spaces that the user creates in the domain. SageMaker doesn't apply these settings to shared spaces.

Type: Array of CustomFileSystemConfig objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

CustomPosixUserConfig

Details about the POSIX identity that is used for file system operations.

SageMaker applies these settings only to private spaces that the user creates in the domain. SageMaker doesn't apply these settings to shared spaces.

Type: CustomPosixUserConfig object

Required: No

DefaultLandingUri

The default experience that the user is directed to when accessing the domain. The supported values are:

- `studio::` Indicates that Studio is the default experience. This value can only be passed if `StudioWebPortal` is set to `ENABLED`.
- `app:JupyterServer:` Indicates that Studio Classic is the default experience.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1023.

Required: No

ExecutionRole

The execution role for the user.

SageMaker applies this setting only to private spaces that the user creates in the domain. SageMaker doesn't apply this setting to shared spaces.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:iam::\d{12}:role/?[a-zA-Z_0-9+=,.@\-_/]+`

Required: No

JupyterLabAppSettings

The settings for the JupyterLab application.

SageMaker applies these settings only to private spaces that the user creates in the domain. SageMaker doesn't apply these settings to shared spaces.

Type: JupyterLabAppSettings object

Required: No

JupyterServerAppSettings

The Jupyter server's app settings.

Type: JupyterServerAppSettings object

Required: No

KernelGatewayAppSettings

The kernel gateway app settings.

Type: KernelGatewayAppSettings object

Required: No

RSessionAppSettings

A collection of settings that configure the `RSessionGateway` app.

Type: RSessionAppSettings object

Required: No

RStudioServerProAppSettings

A collection of settings that configure user interaction with the `RStudioServerPro` app.

Type: RStudioServerProAppSettings object

Required: No

SecurityGroups

The security groups for the Amazon Virtual Private Cloud (VPC) that the domain uses for communication.

Optional when the `CreateDomain.AppNetworkAccessType` parameter is set to `PublicInternetOnly`.

Required when the `CreateDomain.AppNetworkAccessType` parameter is set to `VpcOnly`, unless specified as part of the `DefaultUserSettings` for the domain.

Amazon SageMaker AI adds a security group to allow NFS traffic from Amazon SageMaker AI Studio. Therefore, the number of security groups that you can specify is one less than the maximum number shown.

SageMaker applies these settings only to private spaces that the user creates in the domain. SageMaker doesn't apply these settings to shared spaces.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: No

SharingSettings

Specifies options for sharing Amazon SageMaker AI Studio notebooks.

Type: SharingSettings object

Required: No

SpaceStorageSettings

The storage settings for a space.

SageMaker applies these settings only to private spaces that the user creates in the domain. SageMaker doesn't apply these settings to shared spaces.

Type: DefaultSpaceStorageSettings object

Required: No

StudioWebPortal

Whether the user can access Studio. If this value is set to `DISABLED`, the user cannot access Studio, even if that is the default experience for the domain.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

StudioWebPortalSettings

Studio settings. If these settings are applied on a user level, they take priority over the settings applied on a domain level.

Type: StudioWebPortalSettings object

Required: No

TensorBoardAppSettings

The TensorBoard app settings.

Type: TensorBoardAppSettings object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VariantProperty

Service: Amazon SageMaker Service

Specifies a production variant property type for an Endpoint.

If you are updating an endpoint with the `RetainAllVariantProperties` option of UpdateEndpointInput set to `true`, the `VariantProperty` objects listed in the `ExcludeRetainedVariantProperties` parameter of UpdateEndpointInput override the existing variant properties of the endpoint.

Contents

VariantPropertyType

The type of variant property. The supported values are:

- `DesiredInstanceCount`: Overrides the existing variant instance counts using the `InitialInstanceCount` values in the `ProductionVariants` of CreateEndpointConfig.
- `DesiredWeight`: Overrides the existing variant weights using the `InitialVariantWeight` values in the `ProductionVariants` of CreateEndpointConfig.
- `DataCaptureConfig`: (Not currently supported.)

Type: String

Valid Values: `DesiredInstanceCount | DesiredWeight | DataCaptureConfig`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorConfig

Service: Amazon SageMaker Service

Configuration for your vector collection type.

Contents

Dimension

The number of elements in your vector.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 8192.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Vertex

Service: Amazon SageMaker Service

A lineage entity connected to the starting entity(ies).

Contents

Arn

The Amazon Resource Name (ARN) of the lineage entity resource.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:(experiment|experiment-trial-component|artifact|action|context)/.*`

Required: No

LineageType

The type of resource of the lineage entity.

Type: String

Valid Values: `TrialComponent | Artifact | Context | Action`

Required: No

Type

The type of the lineage entity resource. For example: `DataSet`, `Model`, `Endpoint`, etc...

Type: String

Length Constraints: Minimum length of 0. Maximum length of 40.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VisibilityConditions

Service: Amazon SageMaker Service

The list of key-value pairs used to filter your search results. If a search result contains a key from your list, it is included in the final search response if the value associated with the key in the result matches the value you specified. If the value doesn't match, the result is excluded from the search response. Any resources that don't have a key from the list that you've provided will also be included in the search response.

Contents

Key

The key that specifies the tag that you're using to filter the search results. It must be in the following format: `Tags.<key>`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `([\p{L}\p{Z}\p{N}_.:/=+\-@]*)`

Required: No

Value

The value for the tag that you're using to filter the search results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `([\p{L}\p{Z}\p{N}_.:/=+\-@]*)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

VpcConfig

Service: Amazon SageMaker Service

Specifies an Amazon Virtual Private Cloud (VPC) that your SageMaker jobs, hosted models, and compute resources have access to. You can control access to and from your resources by configuring a VPC. For more information, see Give SageMaker Access to Resources in your Amazon VPC.

Contents

SecurityGroupIds

The VPC security group IDs, in the form `sg-xxxxxxxx`. Specify the security groups for the VPC that is specified in the `Subnets` field.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: Yes

Subnets

The ID of the subnets in the VPC to which you want to connect your training job or model. For information about the availability of specific instance types, see Supported Instance Types and Availability Zones.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

WarmPoolStatus

Service: Amazon SageMaker Service

Status and billing information about the warm pool.

Contents

Status

The status of the warm pool.

- `InUse`: The warm pool is in use for the training job.
- `Available`: The warm pool is available to reuse for a matching training job.
- `Reused`: The warm pool moved to a matching training job for reuse.
- `Terminated`: The warm pool is no longer available. Warm pools are unavailable if they are terminated by a user, terminated for a patch update, or terminated for exceeding the specified `KeepAlivePeriodInSeconds`.

Type: String

Valid Values: `Available | Terminated | Reused | InUse`

Required: Yes

ResourceRetainedBillableTimeInSeconds

The billable time in seconds used by the warm pool. Billable time refers to the absolute wall-clock time.

Multiply `ResourceRetainedBillableTimeInSeconds` by the number of instances (`InstanceCount`) in your training cluster to get the total compute time SageMaker bills you if you run warm pool training. The formula is as follows: `ResourceRetainedBillableTimeInSeconds * InstanceCount`.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

ReusedByJob

The name of the matching training job that reused the warm pool.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

WorkerAccessConfiguration

Service: Amazon SageMaker Service

Use this optional parameter to constrain access to an Amazon S3 resource based on the IP address using supported IAM global condition keys. The Amazon S3 resource is accessed in the worker portal using a Amazon S3 presigned URL.

Contents

S3Presign

Defines any Amazon S3 resource constraints.

Type: S3Presign object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Workforce

Service: Amazon SageMaker Service

A single private workforce, which is automatically created when you create your first private work team. You can create one private work force in each AWS Region. By default, any workforce-related API operation used in a specific region will apply to the workforce created in that region. To learn how to create a private workforce, see Create a Private Workforce.

Contents

WorkforceArn

The Amazon Resource Name (ARN) of the private workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:workforce/.*`

Required: Yes

WorkforceName

The name of the private workforce.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9\-]){0,62}`

Required: Yes

CognitoConfig

The configuration of an Amazon Cognito workforce. A single Cognito workforce is created using and corresponds to a single Amazon Cognito user pool.

Type: CognitoConfig object

Required: No

CreateDate

The date that the workforce is created.

Type: Timestamp

Required: No

FailureReason

The reason your workforce failed.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `.+`

Required: No

IpAddressType

The IP address type you specify - either `IPv4` only or `dualstack` (`IPv4` and `IPv6`) - to support your labeling workforce.

Type: String

Valid Values: `ipv4 | dualstack`

Required: No

LastUpdatedDate

The most recent date that UpdateWorkforce was used to successfully add one or more IP address ranges (CIDRs) to a private workforce's allow list.

Type: Timestamp

Required: No

OidcConfig

The configuration of an OIDC Identity Provider (IdP) private workforce.

Type: OidcConfigForResponse object

Required: No

SourceIpConfig

A list of one to ten IP address ranges (CIDRs) to be added to the workforce allow list. By default, a workforce isn't restricted to specific IP addresses.

Type: SourceIpConfig object

Required: No

Status

The status of your workforce.

Type: String

Valid Values: `Initializing | Updating | Deleting | Failed | Active`

Required: No

SubDomain

The subdomain for your OIDC Identity Provider.

Type: String

Required: No

WorkforceVpcConfig

The configuration of a VPC workforce.

Type: WorkforceVpcConfigResponse object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

WorkforceVpcConfigRequest

Service: Amazon SageMaker Service

The VPC object you use to create or update a workforce.

Contents

SecurityGroupIds

The VPC security group IDs, in the form `sg-xxxxxxxx`. The security groups must be for the same VPC as specified in the subnet.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `sg-[0-9a-z]*`

Required: No

Subnets

The ID of the subnets in the VPC that you want to connect.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `subnet-[0-9a-z]*`

Required: No

VpcId

The ID of the VPC that the workforce uses for communication.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `vpc-[0-9a-z]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

WorkforceVpcConfigResponse

Service: Amazon SageMaker Service

A VpcConfig object that specifies the VPC that you want your workforce to connect to.

Contents

SecurityGroupIds

The VPC security group IDs, in the form sg-xxxxxxxx. The security groups must be for the same VPC as specified in the subnet.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `sg-[0-9a-z]*`

Required: Yes

Subnets

The ID of the subnets in the VPC that you want to connect.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `subnet-[0-9a-z]*`

Required: Yes

VpcId

The ID of the VPC that the workforce uses for communication.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `vpc-[0-9a-z]*`

Required: Yes

VpcEndpointId

The IDs for the VPC service endpoints of your VPC workforce when it is created and updated.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `vpce-[0-9a-z]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

WorkloadSpec

Service: Amazon SageMaker Service

The workload specification for benchmark tool configuration. Provide an inline YAML or JSON string.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

Inline

An inline YAML or JSON string that defines benchmark parameters.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

WorkspaceSettings

Service: Amazon SageMaker Service

The workspace settings for the SageMaker Canvas application.

Contents

S3ArtifactPath

The Amazon S3 bucket used to store artifacts generated by Canvas. Updating the Amazon S3 location impacts existing configuration settings, and Canvas users no longer have access to their artifacts. Canvas users must log out and log back in to apply the new location.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `(https|s3)://([^/]+)/?(.*)`

Required: No

S3KmsKeyId

The AWS Key Management Service (KMS) encryption key ID that is used to encrypt artifacts generated by Canvas in the Amazon S3 bucket.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9:/_-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

Workteam

Service: Amazon SageMaker Service

Provides details about a labeling work team.

Contents

Description

A description of the work team.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `.+`

Required: Yes

MemberDefinitions

A list of `MemberDefinition` objects that contains objects that identify the workers that make up the work team.

Workforces can be created using Amazon Cognito or your own OIDC Identity Provider (IdP). For private workforces created using Amazon Cognito use `CognitoMemberDefinition`. For workforces created using your own OIDC identity provider (IdP) use `OidcMemberDefinition`.

Type: Array of MemberDefinition objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: Yes

WorkteamArn

The Amazon Resource Name (ARN) that identifies the work team.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:workteam/.*`

Required: Yes

WorkteamName

The name of the work team.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,62}`

Required: Yes

CreateDate

The date and time that the work team was created (timestamp).

Type: Timestamp

Required: No

LastUpdatedDate

The date and time that the work team was last updated (timestamp).

Type: Timestamp

Required: No

NotificationConfiguration

Configures SNS notifications of available or expiring work items for work teams.

Type: NotificationConfiguration object

Required: No

ProductListingIds

The Amazon Marketplace identifier for a vendor's work team.

Type: Array of strings

Required: No

SubDomain

The URI of the labeling job's user interface. Workers open this URI to start labeling your data objects.

Type: String

Required: No

WorkerAccessConfiguration

Describes any access constraints that have been defined for Amazon S3 resources.

Type: WorkerAccessConfiguration object

Required: No

WorkforceArn

The Amazon Resource Name (ARN) of the workforce.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:workforce/.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Amazon SageMaker Runtime

The following data types are supported by Amazon SageMaker Runtime:

- PayloadPart
- ResponseStream

PayloadPart

Service: Amazon SageMaker Runtime

A wrapper for pieces of the payload that's returned in response to a streaming inference request. A streaming inference response consists of one or more payload parts.

Contents

Bytes

A blob that contains part of the response for your streaming inference request.

Type: Base64-encoded binary data object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResponseStream

Service: Amazon SageMaker Runtime

A stream of payload parts. Each part contains a portion of the response for a streaming inference request.

Contents

InternalStreamFailure

The stream processing failed because of an unknown error, exception or failure. Try your request again.

Type: Exception

HTTP Status Code:

Required: No

ModelStreamError

An error occurred while streaming the response body. This error can have the following error codes:

ModelInvocationTimeExceeded

The model failed to finish sending the response within the timeout period allowed by Amazon SageMaker AI.

StreamBroken

The Transmission Control Protocol (TCP) connection between the client and the model was reset or closed.

Type: Exception

HTTP Status Code:

Required: No

PayloadPart

A wrapper for pieces of the payload that's returned in response to a streaming inference request. A streaming inference response consists of one or more payload parts.

Type: PayloadPart object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Amazon Sagemaker Edge Manager

The following data types are supported by Amazon Sagemaker Edge Manager:

- Checksum
- Definition
- DeploymentModel
- DeploymentResult
- EdgeDeployment
- EdgeMetric
- Model

Checksum

Service: Amazon Sagemaker Edge Manager

Information about the checksum of a model deployed on a device.

Contents

Sum

The checksum of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-z0-9](-*[a-z0-9])*$`

Required: No

Type

The type of the checksum.

Type: String

Valid Values: SHA1

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Definition

Service: Amazon Sagemaker Edge Manager

Information about deployment details based on the deployment type.

Contents

Checksum

The checksum information of the model.

Type: Checksum object

Required: No

ModelHandle

The unique model handle.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

S3Url

The absolute S3 location of the model.

Type: String

Length Constraints: Maximum length of 1024.

Pattern: `^s3://([^/]+)/?(.*)$`

Required: No

State

The desired state of the model.

Type: String

Valid Values: DEPLOY | UNDEPLOY

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeploymentModel

Service: Amazon Sagemaker Edge Manager

Information about a model deployed on an edge device that is registered with SageMaker Edge Manager.

Contents

DesiredState

The desired state of the model.

Type: String

Valid Values: `DEPLOY | UNDEPLOY`

Required: No

ModelHandle

The name by which the device refers to the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

ModelName

The name of the model.

Type: String

Length Constraints: Minimum length of 4. Maximum length of 255.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

ModelVersion

The version of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9\ \_\.]+`

Required: No

RollbackFailureReason

Returns the error message if there is a rollback.

Type: String

Required: No

State

Returns the current state of the model.

Type: String

Valid Values: `DEPLOY | UNDEPLOY`

Required: No

Status

Returns the deployment status of the model.

Type: String

Valid Values: `SUCCESS | FAIL`

Required: No

StatusReason

Returns the error message for the deployment status result.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeploymentResult

Service: Amazon Sagemaker Edge Manager

Information about the result of a deployment on an edge device that is registered with SageMaker Edge Manager.

Contents

DeploymentEndTime

The timestamp of when the deployment was ended, and the agent got the deployment results.

Type: Timestamp

Required: No

DeploymentModels

Returns a list of models deployed on the agent.

Type: Array of DeploymentModel objects

Required: No

DeploymentName

The name and unique ID of the deployment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

DeploymentStartTime

The timestamp of when the deployment was started on the agent.

Type: Timestamp

Required: No

DeploymentStatus

Returns the bucket error code.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

DeploymentStatusMessage

Returns the detailed error message.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EdgeDeployment

Service: Amazon Sagemaker Edge Manager

Information about a deployment on an edge device that is registered with SageMaker Edge Manager.

Contents

Definitions

Returns a list of Definition objects.

Type: Array of Definition objects

Required: No

DeploymentName

The name and unique ID of the deployment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

FailureHandlingPolicy

Determines whether to rollback to previous configuration if deployment fails.

Type: String

Valid Values: `ROLLBACK_ON_FAILURE | DO_NOTHING`

Required: No

Type

The type of the deployment.

Type: String

Valid Values: `Model`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EdgeMetric

Service: Amazon Sagemaker Edge Manager

Information required for edge device metrics.

Contents

Dimension

The dimension of metrics published.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1000.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9\/])*$`

Required: No

MetricName

Returns the name of the metric.

Type: String

Length Constraints: Minimum length of 4. Maximum length of 100.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

Timestamp

Timestamp of when the metric was requested.

Type: Timestamp

Required: No

Value

Returns the value of the metric.

Type: Double

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Model

Service: Amazon Sagemaker Edge Manager

Information about a model deployed on an edge device that is registered with SageMaker Edge Manager.

Contents

LatestInference

The timestamp of the last inference that was made.

Type: Timestamp

Required: No

LatestSampleTime

The timestamp of the last data sample taken.

Type: Timestamp

Required: No

ModelMetrics

Information required for model metrics.

Type: Array of EdgeMetric objects

Required: No

ModelName

The name of the model.

Type: String

Length Constraints: Minimum length of 4. Maximum length of 255.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

ModelVersion

The version of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9\ \_\.]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Amazon SageMaker Feature Store Runtime

The following data types are supported by Amazon SageMaker Feature Store Runtime:

- BatchGetRecordError
- BatchGetRecordIdentifier
- BatchGetRecordResultDetail
- FeatureValue
- TtlDuration

BatchGetRecordError

Service: Amazon SageMaker Feature Store Runtime

The error that has occurred when attempting to retrieve a batch of Records.

Contents

ErrorCode

The error code of an error that has occurred when attempting to retrieve a batch of Records. For more information on errors, see Errors.

Type: String

Length Constraints: Maximum length of 358400.

Pattern: `.*`

Required: Yes

ErrorMessage

The error message of an error that has occurred when attempting to retrieve a record in the batch.

Type: String

Length Constraints: Maximum length of 2048.

Required: Yes

FeatureGroupName

The name of the feature group that the record belongs to.

Type: String

Length Constraints: Maximum length of 358400.

Pattern: `.*`

Required: Yes

RecordIdentifierValueAsString

The value for the `RecordIdentifier` in string format of a Record from a `FeatureGroup` that is causing an error when attempting to be retrieved.

Type: String

Length Constraints: Maximum length of 358400.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchGetRecordIdentifier

Service: Amazon SageMaker Feature Store Runtime

The identifier that identifies the batch of Records you are retrieving in a batch.

Contents

FeatureGroupName

The name or Amazon Resource Name (ARN) of the `FeatureGroup` containing the records you are retrieving in a batch.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 150.

Pattern: `(arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:feature-group/)?([a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63})`

Required: Yes

RecordIdentifiersValueAsString

The value for a list of record identifiers in string format.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Length Constraints: Maximum length of 358400.

Pattern: `.*`

Required: Yes

FeatureNames

List of names of Features to be retrieved. If not specified, the latest value for all the Features are returned.

Type: Array of strings

Array Members: Minimum number of 1 item.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `^[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchGetRecordResultDetail

Service: Amazon SageMaker Feature Store Runtime

The output of records that have been retrieved in a batch.

Contents

FeatureGroupName

The `FeatureGroupName` containing Records you retrieved in a batch.

Type: String

Length Constraints: Maximum length of 358400.

Pattern: `.*`

Required: Yes

Record

The `Record` retrieved.

Type: Array of FeatureValue objects

Array Members: Minimum number of 1 item.

Required: Yes

RecordIdentifierValueAsString

The value of the record identifier in string format.

Type: String

Length Constraints: Maximum length of 358400.

Pattern: `.*`

Required: Yes

ExpiresAt

The `ExpiresAt` ISO string of the requested record.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FeatureValue

Service: Amazon SageMaker Feature Store Runtime

The value associated with a feature.

Contents

FeatureName

The name of a feature that a feature value corresponds to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `^[a-zA-Z0-9]([-_]*[a-zA-Z0-9]){0,63}`

Required: Yes

ValueAsString

The value in string format associated with a feature. Used when your `CollectionType` is `None`. Note that features types can be `String`, `Integral`, or `Fractional`. This value represents all three types as a string.

Type: String

Length Constraints: Maximum length of 358400.

Pattern: `.*`

Required: No

ValueAsStringList

The list of values in string format associated with a feature. Used when your `CollectionType` is a `List`, `Set`, or `Vector`. Note that features types can be `String`, `Integral`, or `Fractional`. These values represents all three types as a string.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 358400 items.

Length Constraints: Maximum length of 358400.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TtlDuration

Service: Amazon SageMaker Feature Store Runtime

Time to live duration, where the record is hard deleted after the expiration time is reached; `ExpiresAt` = `EventTime` + `TtlDuration`. For information on HardDelete, see the DeleteRecord API in the Amazon SageMaker API Reference guide.

Contents

Unit

`TtlDuration` time unit.

Type: String

Valid Values: `Seconds | Minutes | Hours | Days | Weeks`

Required: Yes

Value

`TtlDuration` time value.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Amazon SageMaker geospatial capabilities

The following data types are supported by Amazon SageMaker geospatial capabilities:

- AreaOfInterest
- AreaOfInterestGeometry
- AssetValue
- BandMathConfigInput
- CloudMaskingConfigInput
- CloudRemovalConfigInput
- CustomIndicesInput
- EarthObservationJobErrorDetails
- EoCloudCoverInput
- ExportErrorDetails
- ExportErrorDetailsOutput
- ExportS3DataInput
- ExportVectorEnrichmentJobOutputConfig
- Filter
- Geometry
- GeoMosaicConfigInput
- InputConfigInput
- InputConfigOutput
- ItemSource
- JobConfigInput
- LandCoverSegmentationConfigInput
- LandsatCloudCoverLandInput
- ListEarthObservationJobOutputConfig
- ListVectorEnrichmentJobOutputConfig
- MapMatchingConfig
- MultiPolygonGeometryInput
- Operation
- OutputBand
- OutputConfigInput
- OutputResolutionResamplingInput

- OutputResolutionStackInput
- PlatformInput
- PolygonGeometryInput
- Properties
- Property
- PropertyFilter
- PropertyFilters
- RasterDataCollectionMetadata
- RasterDataCollectionQueryInput
- RasterDataCollectionQueryOutput
- RasterDataCollectionQueryWithBandFilterInput
- ResamplingConfigInput
- ReverseGeocodingConfig
- StackConfigInput
- TemporalStatisticsConfigInput
- TimeRangeFilterInput
- TimeRangeFilterOutput
- UserDefined
- VectorEnrichmentJobConfig
- VectorEnrichmentJobDataSourceConfigInput
- VectorEnrichmentJobErrorDetails
- VectorEnrichmentJobExportErrorDetails
- VectorEnrichmentJobInputConfig
- VectorEnrichmentJobS3Data
- ViewOffNadirInput
- ViewSunAzimuthInput
- ViewSunElevationInput
- ZonalStatisticsConfigInput

AreaOfInterest

Service: Amazon SageMaker geospatial capabilities

The geographic extent of the Earth Observation job.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

AreaOfInterestGeometry

A GeoJSON object representing the geographic extent in the coordinate space.

Type: AreaOfInterestGeometry object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AreaOfInterestGeometry

Service: Amazon SageMaker geospatial capabilities

A GeoJSON object representing the geographic extent in the coordinate space.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

MultiPolygonGeometry

The structure representing the MultiPolygon Geometry.

Type: MultiPolygonGeometryInput object

Required: No

PolygonGeometry

The structure representing Polygon Geometry.

Type: PolygonGeometryInput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AssetValue

Service: Amazon SageMaker geospatial capabilities

The structure containing the asset properties.

Contents

Href

Link to the asset object.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BandMathConfigInput

Service: Amazon SageMaker geospatial capabilities

Input structure for the BandMath operation type. Defines Predefined and CustomIndices to be computed using BandMath.

Contents

CustomIndices

CustomIndices that are computed.

Type: CustomIndicesInput object

Required: No

PredefinedIndices

One or many of the supported predefined indices to compute. Allowed values: `NDVI`, `EVI2`, `MSAVI`, `NDWI`, `NDMI`, `NDSI`, and `WDRVI`.

Type: Array of strings

Array Members: Minimum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CloudMaskingConfigInput

Service: Amazon SageMaker geospatial capabilities

Input structure for CloudMasking operation type.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CloudRemovalConfigInput

Service: Amazon SageMaker geospatial capabilities

Input structure for Cloud Removal Operation type

Contents

AlgorithmName

The name of the algorithm used for cloud removal.

Type: String

Valid Values: `INTERPOLATION`

Required: No

InterpolationValue

The interpolation value you provide for cloud removal.

Type: String

Required: No

TargetBands

TargetBands to be returned in the output of CloudRemoval operation.

Type: Array of strings

Array Members: Minimum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomIndicesInput

Service: Amazon SageMaker geospatial capabilities

Input object defining the custom BandMath indices to compute.

Contents

Operations

A list of BandMath indices to compute.

Type: Array of Operation objects

Array Members: Minimum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EarthObservationJobErrorDetails

Service: Amazon SageMaker geospatial capabilities

The structure representing the errors in an EarthObservationJob.

Contents

Message

A detailed message describing the error in an Earth Observation job.

Type: String

Required: No

Type

The type of error in an Earth Observation job.

Type: String

Valid Values: `CLIENT_ERROR | SERVER_ERROR`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EoCloudCoverInput

Service: Amazon SageMaker geospatial capabilities

The structure representing the EoCloudCover filter.

Contents

LowerBound

Lower bound for EoCloudCover.

Type: Float

Required: Yes

UpperBound

Upper bound for EoCloudCover.

Type: Float

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExportErrorDetails

Service: Amazon SageMaker geospatial capabilities

The structure for returning the export error details in a GetEarthObservationJob.

Contents

ExportResults

The structure for returning the export error details while exporting results of an Earth Observation job.

Type: ExportErrorDetailsOutput object

Required: No

ExportSourceImages

The structure for returning the export error details while exporting the source images of an Earth Observation job.

Type: ExportErrorDetailsOutput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExportErrorDetailsOutput

Service: Amazon SageMaker geospatial capabilities

The structure representing the errors in an export EarthObservationJob operation.

Contents

Message

A detailed message describing the error in an export EarthObservationJob operation.

Type: String

Required: No

Type

The type of error in an export EarthObservationJob operation.

Type: String

Valid Values: `CLIENT_ERROR | SERVER_ERROR`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExportS3DataInput

Service: Amazon SageMaker geospatial capabilities

The structure containing the Amazon S3 path to export the Earth Observation job output.

Contents

S3Uri

The URL to the Amazon S3 data input.

Type: String

Pattern: `s3://([^/]+)/?(.*)`

Required: Yes

KmsKeyId

The Key Management Service key ID for server-side encryption.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExportVectorEnrichmentJobOutputConfig

Service: Amazon SageMaker geospatial capabilities

An object containing information about the output file.

Contents

S3Data

The input structure for Amazon S3 data; representing the Amazon S3 location of the input data objects.

Type: VectorEnrichmentJobS3Data object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Filter

Service: Amazon SageMaker geospatial capabilities

The structure representing the filters supported by a RasterDataCollection.

Contents

Name

The name of the filter.

Type: String

Required: Yes

Type

The type of the filter being used.

Type: String

Required: Yes

Maximum

The maximum value of the filter.

Type: Float

Required: No

Minimum

The minimum value of the filter.

Type: Float

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Geometry

Service: Amazon SageMaker geospatial capabilities

The structure representing a Geometry in terms of Type and Coordinates as per GeoJson spec.

Contents

Coordinates

The coordinates of the GeoJson Geometry.

Type: Array of arrays of arrays of doubles

Array Members: Minimum number of 1 item.

Array Members: Minimum number of 4 items.

Array Members: Fixed number of 2 items.

Required: Yes

Type

GeoJson Geometry types like Polygon and MultiPolygon.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GeoMosaicConfigInput

Service: Amazon SageMaker geospatial capabilities

Input configuration information for the geomosaic.

Contents

AlgorithmName

The name of the algorithm being used for geomosaic.

Type: String

Valid Values: `NEAR | BILINEAR | CUBIC | CUBICSPLINE | LANCZOS | AVERAGE | RMS | MODE | MAX | MIN | MED | Q1 | Q3 | SUM`

Required: No

TargetBands

The target bands for geomosaic.

Type: Array of strings

Array Members: Minimum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InputConfigInput

Service: Amazon SageMaker geospatial capabilities

Input configuration information.

Contents

PreviousEarthObservationJobArn

The Amazon Resource Name (ARN) of the previous Earth Observation job.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:earth-observation-job/[a-z0-9]{12,}`

Required: No

RasterDataCollectionQuery

The structure representing the RasterDataCollection Query consisting of the Area of Interest, RasterDataCollectionArn,TimeRange and Property Filters.

Type: RasterDataCollectionQueryInput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InputConfigOutput

Service: Amazon SageMaker geospatial capabilities

The InputConfig for an EarthObservationJob response.

Contents

PreviousEarthObservationJobArn

The Amazon Resource Name (ARN) of the previous Earth Observation job.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:earth-observation-job/[a-z0-9]{12,}`

Required: No

RasterDataCollectionQuery

The structure representing the RasterDataCollection Query consisting of the Area of Interest, RasterDataCollectionArn, RasterDataCollectionName, TimeRange, and Property Filters.

Type: RasterDataCollectionQueryOutput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ItemSource

Service: Amazon SageMaker geospatial capabilities

The structure representing the items in the response for SearchRasterDataCollection.

Contents

DateTime

The searchable date and time of the item, in UTC.

Type: Timestamp

Required: Yes

Geometry

The item Geometry in GeoJson format.

Type: Geometry object

Required: Yes

Id

A unique Id for the source item.

Type: String

Required: Yes

Assets

This is a dictionary of Asset Objects data associated with the Item that can be downloaded or streamed, each with a unique key.

Type: String to AssetValue object map

Required: No

Properties

This field contains additional properties of the item.

Type: Properties object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

JobConfigInput

Service: Amazon SageMaker geospatial capabilities

The input structure for the JobConfig in an EarthObservationJob.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

BandMathConfig

An object containing information about the job configuration for BandMath.

Type: BandMathConfigInput object

Required: No

CloudMaskingConfig

An object containing information about the job configuration for cloud masking.

Type: CloudMaskingConfigInput object

Required: No

CloudRemovalConfig

An object containing information about the job configuration for cloud removal.

Type: CloudRemovalConfigInput object

Required: No

GeoMosaicConfig

An object containing information about the job configuration for geomosaic.

Type: GeoMosaicConfigInput object

Required: No

LandCoverSegmentationConfig

An object containing information about the job configuration for land cover segmentation.

Type: LandCoverSegmentationConfigInput object

Required: No

ResamplingConfig

An object containing information about the job configuration for resampling.

Type: ResamplingConfigInput object

Required: No

StackConfig

An object containing information about the job configuration for a Stacking Earth Observation job.

Type: StackConfigInput object

Required: No

TemporalStatisticsConfig

An object containing information about the job configuration for temporal statistics.

Type: TemporalStatisticsConfigInput object

Required: No

ZonalStatisticsConfig

An object containing information about the job configuration for zonal statistics.

Type: ZonalStatisticsConfigInput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LandCoverSegmentationConfigInput

Service: Amazon SageMaker geospatial capabilities

The input structure for Land Cover Operation type.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LandsatCloudCoverLandInput

Service: Amazon SageMaker geospatial capabilities

The structure representing Land Cloud Cover property for Landsat data collection.

Contents

LowerBound

The minimum value for Land Cloud Cover property filter. This will filter items having Land Cloud Cover greater than or equal to this value.

Type: Float

Required: Yes

UpperBound

The maximum value for Land Cloud Cover property filter. This will filter items having Land Cloud Cover less than or equal to this value.

Type: Float

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ListEarthObservationJobOutputConfig

Service: Amazon SageMaker geospatial capabilities

An object containing information about the output file.

Contents

Arn

The Amazon Resource Name (ARN) of the list of the Earth Observation jobs.

Type: String

Required: Yes

CreationTime

The creation time.

Type: Timestamp

Required: Yes

DurationInSeconds

The duration of the session, in seconds.

Type: Integer

Required: Yes

Name

The names of the Earth Observation jobs in the list.

Type: String

Required: Yes

OperationType

The operation type for an Earth Observation job.

Type: String

Required: Yes

Status

The status of the list of the Earth Observation jobs.

Type: String

Valid Values: `INITIALIZING | IN_PROGRESS | STOPPING | COMPLETED | STOPPED | FAILED | DELETING | DELETED`

Required: Yes

Tags

Each tag consists of a key and a value.

Type: String to string map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ListVectorEnrichmentJobOutputConfig

Service: Amazon SageMaker geospatial capabilities

An object containing information about the output file.

Contents

Arn

The Amazon Resource Name (ARN) of the list of the Vector Enrichment jobs.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:vector-enrichment-job/[a-z0-9]{12,}`

Required: Yes

CreationTime

The creation time.

Type: Timestamp

Required: Yes

DurationInSeconds

The duration of the session, in seconds.

Type: Integer

Required: Yes

Name

The names of the Vector Enrichment jobs in the list.

Type: String

Required: Yes

Status

The status of the Vector Enrichment jobs list.

Type: String

Valid Values: `INITIALIZING | IN_PROGRESS | STOPPING | STOPPED | COMPLETED | FAILED | DELETING | DELETED`

Required: Yes

Type

The type of the list of Vector Enrichment jobs.

Type: String

Valid Values: `REVERSE_GEOCODING | MAP_MATCHING`

Required: Yes

Tags

Each tag consists of a key and a value.

Type: String to string map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MapMatchingConfig

Service: Amazon SageMaker geospatial capabilities

The input structure for Map Matching operation type.

Contents

IdAttributeName

The field name for the data that describes the identifier representing a collection of GPS points belonging to an individual trace.

Type: String

Required: Yes

TimestampAttributeName

The name of the timestamp attribute.

Type: String

Required: Yes

XAttributeName

The name of the X-attribute

Type: String

Required: Yes

YAttributeName

The name of the Y-attribute

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MultiPolygonGeometryInput

Service: Amazon SageMaker geospatial capabilities

The structure representing Polygon Geometry based on the GeoJson spec.

Contents

Coordinates

The coordinates of the multipolygon geometry.

Type: Array of arrays of arrays of arrays of doubles

Array Members: Minimum number of 1 item.

Array Members: Minimum number of 4 items.

Array Members: Fixed number of 2 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Operation

Service: Amazon SageMaker geospatial capabilities

Represents an arithmetic operation to compute spectral index.

Contents

Equation

Textual representation of the math operation; Equation used to compute the spectral index.

Type: String

Required: Yes

Name

The name of the operation.

Type: String

Required: Yes

OutputType

The type of the operation.

Type: String

Valid Values: `INT32 | FLOAT32 | INT16 | FLOAT64 | UINT16`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputBand

Service: Amazon SageMaker geospatial capabilities

A single EarthObservationJob output band.

Contents

BandName

The name of the band.

Type: String

Required: Yes

OutputDataType

The datatype of the output band.

Type: String

Valid Values: `INT32 | FLOAT32 | INT16 | FLOAT64 | UINT16`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputConfigInput

Service: Amazon SageMaker geospatial capabilities

The response structure for an OutputConfig returned by an ExportEarthObservationJob.

Contents

S3Data

Path to Amazon S3 storage location for the output configuration file.

Type: ExportS3DataInput object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputResolutionResamplingInput

Service: Amazon SageMaker geospatial capabilities

OutputResolution Configuration indicating the target resolution for the output of Resampling operation.

Contents

UserDefined

User Defined Resolution for the output of Resampling operation defined by value and unit.

Type: UserDefined object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputResolutionStackInput

Service: Amazon SageMaker geospatial capabilities

The input structure representing Output Resolution for Stacking Operation.

Contents

Predefined

A string value representing Predefined Output Resolution for a stacking operation. Allowed values are `HIGHEST`, `LOWEST`, and `AVERAGE`.

Type: String

Valid Values: `HIGHEST | LOWEST | AVERAGE`

Required: No

UserDefined

The structure representing User Output Resolution for a Stacking operation defined as a value and unit.

Type: UserDefined object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PlatformInput

Service: Amazon SageMaker geospatial capabilities

The input structure for specifying Platform. Platform refers to the unique name of the specific platform the instrument is attached to. For satellites it is the name of the satellite, eg. landsat-8 (Landsat-8), sentinel-2a.

Contents

Value

The value of the platform.

Type: String

Required: Yes

ComparisonOperator

The ComparisonOperator to use with PlatformInput.

Type: String

Valid Values: `EQUALS | NOT_EQUALS | STARTS_WITH`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PolygonGeometryInput

Service: Amazon SageMaker geospatial capabilities

The structure representing Polygon Geometry based on the GeoJson spec.

Contents

Coordinates

Coordinates representing a Polygon based on the GeoJson spec.

Type: Array of arrays of arrays of doubles

Array Members: Minimum number of 1 item.

Array Members: Minimum number of 4 items.

Array Members: Fixed number of 2 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Properties

Service: Amazon SageMaker geospatial capabilities

Properties associated with the Item.

Contents

EoCloudCover

Estimate of cloud cover.

Type: Float

Required: No

LandsatCloudCoverLand

Land cloud cover for Landsat Data Collection.

Type: Float

Required: No

Platform

Platform property. Platform refers to the unique name of the specific platform the instrument is attached to. For satellites it is the name of the satellite, eg. landsat-8 (Landsat-8), sentinel-2a.

Type: String

Required: No

ViewOffNadir

The angle from the sensor between nadir (straight down) and the scene center. Measured in degrees (0-90).

Type: Float

Required: No

ViewSunAzimuth

The sun azimuth angle. From the scene center point on the ground, this is the angle between truth north and the sun. Measured clockwise in degrees (0-360).

Type: Float

Required: No

ViewSunElevation

The sun elevation angle. The angle from the tangent of the scene center point to the sun. Measured from the horizon in degrees (-90-90). Negative values indicate the sun is below the horizon, e.g. sun elevation of -10° means the data was captured during nautical twilight.

Type: Float

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Property

Service: Amazon SageMaker geospatial capabilities

Represents a single searchable property to search on.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

EoCloudCover

The structure representing EoCloudCover property filter containing a lower bound and upper bound.

Type: EoCloudCoverInput object

Required: No

LandsatCloudCoverLand

The structure representing Land Cloud Cover property filter for Landsat collection containing a lower bound and upper bound.

Type: LandsatCloudCoverLandInput object

Required: No

Platform

The structure representing Platform property filter consisting of value and comparison operator.

Type: PlatformInput object

Required: No

ViewOffNadir

The structure representing ViewOffNadir property filter containing a lower bound and upper bound.

Type: ViewOffNadirInput object

Required: No

ViewSunAzimuth

The structure representing ViewSunAzimuth property filter containing a lower bound and upper bound.

Type: ViewSunAzimuthInput object

Required: No

ViewSunElevation

The structure representing ViewSunElevation property filter containing a lower bound and upper bound.

Type: ViewSunElevationInput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PropertyFilter

Service: Amazon SageMaker geospatial capabilities

The structure representing a single PropertyFilter.

Contents

Property

Represents a single property to match with when searching a raster data collection.

Type: Property object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PropertyFilters

Service: Amazon SageMaker geospatial capabilities

A list of PropertyFilter objects.

Contents

LogicalOperator

The Logical Operator used to combine the Property Filters.

Type: String

Valid Values: `AND`

Required: No

Properties

A list of Property Filters.

Type: Array of PropertyFilter objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RasterDataCollectionMetadata

Service: Amazon SageMaker geospatial capabilities

Response object containing details for a specific RasterDataCollection.

Contents

Arn

The Amazon Resource Name (ARN) of the raster data collection.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:raster-data-collection/(public|premium|user)/[a-z0-9]{12,}`

Required: Yes

Description

A description of the raster data collection.

Type: String

Required: Yes

Name

The name of the raster data collection.

Type: String

Required: Yes

SupportedFilters

The list of filters supported by the raster data collection.

Type: Array of Filter objects

Required: Yes

Type

The type of raster data collection.

Type: String

Valid Values: `PUBLIC | PREMIUM | USER`

Required: Yes

DescriptionPageUrl

The description URL of the raster data collection.

Type: String

Required: No

Tags

Each tag consists of a key and a value.

Type: String to string map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RasterDataCollectionQueryInput

Service: Amazon SageMaker geospatial capabilities

The input structure for Raster Data Collection Query containing the Area of Interest, TimeRange Filters, and Property Filters.

Contents

RasterDataCollectionArn

The Amazon Resource Name (ARN) of the raster data collection.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:raster-data-collection/(public|premium|user)/[a-z0-9]{12,}`

Required: Yes

TimeRangeFilter

The TimeRange Filter used in the RasterDataCollection Query.

Type: TimeRangeFilterInput object

Required: Yes

AreaOfInterest

The area of interest being queried for the raster data collection.

Type: AreaOfInterest object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

PropertyFilters

The list of Property filters used in the Raster Data Collection Query.

Type: PropertyFilters object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RasterDataCollectionQueryOutput

Service: Amazon SageMaker geospatial capabilities

The output structure contains the Raster Data Collection Query input along with some additional metadata.

Contents

RasterDataCollectionArn

The ARN of the Raster Data Collection against which the search is done.

Type: String

Pattern: `arn:aws[a-z-]{0,12}:sagemaker-geospatial:[a-z0-9-]{1,25}:[0-9]{12}:raster-data-collection/(public|premium|user)/[a-z0-9]{12,}`

Required: Yes

RasterDataCollectionName

The name of the raster data collection.

Type: String

Required: Yes

TimeRangeFilter

The TimeRange filter used in the search.

Type: TimeRangeFilterOutput object

Required: Yes

AreaOfInterest

The Area of Interest used in the search.

Type: AreaOfInterest object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

PropertyFilters

Property filters used in the search.

Type: PropertyFilters object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RasterDataCollectionQueryWithBandFilterInput

Service: Amazon SageMaker geospatial capabilities

This is a RasterDataCollectionQueryInput containing AreaOfInterest, Time Range filter and Property filters.

Contents

TimeRangeFilter

The TimeRange Filter used in the search query.

Type: TimeRangeFilterInput object

Required: Yes

AreaOfInterest

The Area of interest to be used in the search query.

Type: AreaOfInterest object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

BandFilter

The list of Bands to be displayed in the result for each item.

Type: Array of strings

Array Members: Minimum number of 1 item.

Required: No

PropertyFilters

The Property Filters used in the search query.

Type: PropertyFilters object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResamplingConfigInput

Service: Amazon SageMaker geospatial capabilities

The structure representing input for resampling operation.

Contents

OutputResolution

The structure representing output resolution (in target georeferenced units) of the result of resampling operation.

Type: OutputResolutionResamplingInput object

Required: Yes

AlgorithmName

The name of the algorithm used for resampling.

Type: String

Valid Values: `NEAR | BILINEAR | CUBIC | CUBICSPLINE | LANCZOS | AVERAGE | RMS | MODE | MAX | MIN | MED | Q1 | Q3 | SUM`

Required: No

TargetBands

Bands used in the operation. If no target bands are specified, it uses all bands available in the input.

Type: Array of strings

Array Members: Minimum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ReverseGeocodingConfig

Service: Amazon SageMaker geospatial capabilities

The input structure for Reverse Geocoding operation type.

Contents

XAttributeName

The field name for the data that describes x-axis coordinate, eg. longitude of a point.

Type: String

Required: Yes

YAttributeName

The field name for the data that describes y-axis coordinate, eg. latitude of a point.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StackConfigInput

Service: Amazon SageMaker geospatial capabilities

The input structure for Stacking Operation.

Contents

OutputResolution

The structure representing output resolution (in target georeferenced units) of the result of stacking operation.

Type: OutputResolutionStackInput object

Required: No

TargetBands

A list of bands to be stacked in the specified order. When the parameter is not provided, all the available bands in the data collection are stacked in the alphabetical order of their asset names.

Type: Array of strings

Array Members: Minimum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TemporalStatisticsConfigInput

Service: Amazon SageMaker geospatial capabilities

The structure representing the configuration for Temporal Statistics operation.

Contents

Statistics

The list of the statistics method options.

Type: Array of strings

Array Members: Minimum number of 1 item.

Valid Values: `MEAN | MEDIAN | STANDARD_DEVIATION`

Required: Yes

GroupBy

The input for the temporal statistics grouping by time frequency option.

Type: String

Valid Values: `ALL | YEARLY`

Required: No

TargetBands

The list of target band names for the temporal statistic to calculate.

Type: Array of strings

Array Members: Minimum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TimeRangeFilterInput

Service: Amazon SageMaker geospatial capabilities

The input for the time-range filter.

Contents

EndTime

The end time for the time-range filter.

Type: Timestamp

Required: Yes

StartTime

The start time for the time-range filter.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TimeRangeFilterOutput

Service: Amazon SageMaker geospatial capabilities

The output structure of the time range filter.

Contents

EndTime

The ending time for the time range filter.

Type: Timestamp

Required: Yes

StartTime

The starting time for the time range filter.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UserDefined

Service: Amazon SageMaker geospatial capabilities

The output resolution (in target georeferenced units) of the result of the operation

Contents

Unit

The units for output resolution of the result.

Type: String

Valid Values: `METERS`

Required: Yes

Value

The value for output resolution of the result.

Type: Float

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorEnrichmentJobConfig

Service: Amazon SageMaker geospatial capabilities

It contains configs such as ReverseGeocodingConfig and MapMatchingConfig.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

MapMatchingConfig

The input structure for Map Matching operation type.

Type: MapMatchingConfig object

Required: No

ReverseGeocodingConfig

The input structure for Reverse Geocoding operation type.

Type: ReverseGeocodingConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorEnrichmentJobDataSourceConfigInput

Service: Amazon SageMaker geospatial capabilities

The input structure for the data source that represents the storage type of the input data objects.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

S3Data

The input structure for the Amazon S3 data that represents the Amazon S3 location of the input data objects.

Type: VectorEnrichmentJobS3Data object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorEnrichmentJobErrorDetails

Service: Amazon SageMaker geospatial capabilities

VectorEnrichmentJob error details in response from GetVectorEnrichmentJob.

Contents

ErrorMessage

A message that you define and then is processed and rendered by the Vector Enrichment job when the error occurs.

Type: String

Required: No

ErrorType

The type of error generated during the Vector Enrichment job.

Type: String

Valid Values: `CLIENT_ERROR | SERVER_ERROR`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorEnrichmentJobExportErrorDetails

Service: Amazon SageMaker geospatial capabilities

VectorEnrichmentJob export error details in response from GetVectorEnrichmentJob.

Contents

Message

The message providing details about the errors generated during the Vector Enrichment job.

Type: String

Required: No

Type

The output error details for an Export operation on a Vector Enrichment job.

Type: String

Valid Values: `CLIENT_ERROR | SERVER_ERROR`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorEnrichmentJobInputConfig

Service: Amazon SageMaker geospatial capabilities

The input structure for the InputConfig in a VectorEnrichmentJob.

Contents

DataSourceConfig

The input structure for the data source that represents the storage type of the input data objects.

Type: VectorEnrichmentJobDataSourceConfigInput object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

DocumentType

The input structure that defines the data source file type.

Type: String

Valid Values: CSV

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorEnrichmentJobS3Data

Service: Amazon SageMaker geospatial capabilities

The Amazon S3 data for the Vector Enrichment job.

Contents

S3Uri

The URL to the Amazon S3 data for the Vector Enrichment job.

Type: String

Pattern: `s3://([^/]+)/?(.*)`

Required: Yes

KmsKeyId

The Key Management Service key ID for server-side encryption.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ViewOffNadirInput

Service: Amazon SageMaker geospatial capabilities

The input structure for specifying ViewOffNadir property filter. ViewOffNadir refers to the angle from the sensor between nadir (straight down) and the scene center. Measured in degrees (0-90).

Contents

LowerBound

The minimum value for ViewOffNadir property filter. This filters items having ViewOffNadir greater than or equal to this value.

Type: Float

Required: Yes

UpperBound

The maximum value for ViewOffNadir property filter. This filters items having ViewOffNadir lesser than or equal to this value.

Type: Float

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ViewSunAzimuthInput

Service: Amazon SageMaker geospatial capabilities

The input structure for specifying ViewSunAzimuth property filter. ViewSunAzimuth refers to the Sun azimuth angle. From the scene center point on the ground, this is the angle between truth north and the sun. Measured clockwise in degrees (0-360).

Contents

LowerBound

The minimum value for ViewSunAzimuth property filter. This filters items having ViewSunAzimuth greater than or equal to this value.

Type: Float

Required: Yes

UpperBound

The maximum value for ViewSunAzimuth property filter. This filters items having ViewSunAzimuth lesser than or equal to this value.

Type: Float

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ViewSunElevationInput

Service: Amazon SageMaker geospatial capabilities

The input structure for specifying ViewSunElevation angle property filter.

Contents

LowerBound

The lower bound to view the sun elevation.

Type: Float

Required: Yes

UpperBound

The upper bound to view the sun elevation.

Type: Float

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ZonalStatisticsConfigInput

Service: Amazon SageMaker geospatial capabilities

The structure representing input configuration of ZonalStatistics operation.

Contents

Statistics

List of zonal statistics to compute.

Type: Array of strings

Array Members: Minimum number of 1 item.

Valid Values: `MEAN | MEDIAN | STANDARD_DEVIATION | MAX | MIN | SUM`

Required: Yes

ZoneS3Path

The Amazon S3 path pointing to the GeoJSON containing the polygonal zones.

Type: String

Pattern: `s3://([^/]+)/?(.*)`

Required: Yes

TargetBands

Bands used in the operation. If no target bands are specified, it uses all bands available input.

Type: Array of strings

Array Members: Minimum number of 1 item.

Required: No

ZoneS3PathKmsKeyId

The Amazon Resource Name (ARN) or an ID of a AWS Key Management Service (AWS KMS) key that Amazon SageMaker uses to decrypt your output artifacts with Amazon S3 server-side encryption. The SageMaker execution role must have `kms:GenerateDataKey` permission.

The `KmsKeyId` can be any of the following formats:

- // KMS Key ID

  `"1234abcd-12ab-34cd-56ef-1234567890ab"`

- // Amazon Resource Name (ARN) of a KMS Key

  `"arn:aws:kms:<region>:<account>:key/<key-id-12ab-34cd-56ef-1234567890ab>"`

For more information about key identifiers, see Key identifiers (KeyID) in the AWS Key Management Service (AWS KMS) documentation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Amazon SageMaker Metrics Service

The following data types are supported by Amazon SageMaker Metrics Service:

- BatchPutMetricsError
- MetricQuery
- MetricQueryResult
- RawMetricData

BatchPutMetricsError

Service: Amazon SageMaker Metrics Service

An error that occured when putting the metric data.

Contents

Code

The error code of an error that occured when attempting to put metrics.

- `METRIC_LIMIT_EXCEEDED`: The maximum amount of metrics per resource is exceeded.
- `INTERNAL_ERROR`: An internal error occured.
- `VALIDATION_ERROR`: The metric data failed validation.
- `CONFLICT_ERROR`: Multiple requests attempted to modify the same data simultaneously.

Type: String

Valid Values: `METRIC_LIMIT_EXCEEDED | INTERNAL_ERROR | VALIDATION_ERROR | CONFLICT_ERROR`

Required: No

MetricIndex

An index that corresponds to the metric in the request.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetricQuery

Service: Amazon SageMaker Metrics Service

Specifies a query to retrieve training metrics from SageMaker.

Contents

MetricName

The name of the metric to retrieve.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: Yes

MetricStat

The metrics stat type of metrics to retrieve.

Type: String

Valid Values: `Min | Max | Avg | Count | StdDev | Last`

Required: Yes

Period

The time period of metrics to retrieve.

Type: String

Valid Values: `OneMinute | FiveMinute | OneHour | IterationNumber`

Required: Yes

ResourceArn

The ARN of the SageMaker resource to retrieve metrics for.

Type: String

Length Constraints: Maximum length of 2048.

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]*:[0-9]{12}:[a-z\-].*/.*`

Required: Yes

XAxisType

The x-axis type of metrics to retrieve.

Type: String

Valid Values: `IterationNumber | Timestamp`

Required: Yes

End

The end time of metrics to retrieve.

Type: Long

Required: No

Start

The start time of metrics to retrieve.

Type: Long

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetricQueryResult

Service: Amazon SageMaker Metrics Service

The result of a query to retrieve training metrics from SageMaker.

Contents

MetricValues

The metric values retrieved by the query.

Type: Array of doubles

Required: Yes

Status

The status of the metric query.

Type: String

Valid Values: `Complete | Truncated | InternalError | ValidationError`

Required: Yes

XAxisValues

The values for the x-axis of the metrics.

Type: Array of longs

Required: Yes

Message

A message describing the status of the metric query.

Type: String

Length Constraints: Maximum length of 2048.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RawMetricData

Service: Amazon SageMaker Metrics Service

The raw metric data to associate with the resource.

Contents

MetricName

The name of the metric.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `.+`

Required: Yes

Timestamp

The time that the metric was recorded.

Type: Timestamp

Required: Yes

Value

The metric value.

Type: Double

Required: Yes

Step

The metric step (epoch).

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Amazon SageMaker Job Runtime Service

The following data types are supported by Amazon SageMaker Job Runtime Service:

Common Parameters

The following list contains the parameters that all actions use for signing Signature Version 4 requests with a query string. Any action-specific parameters are listed in the topic for that action. For more information about Signature Version 4, see Signing AWS API requests in the *IAM User Guide*.

X-Amz-Algorithm

The hash algorithm that you used to create the request signature.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Valid Values: `AWS4-HMAC-SHA256`

Required: Conditional

X-Amz-Credential

The credential scope value, which is a string that includes your access key, the date, the region you are targeting, the service you are requesting, and a termination string ("aws4_request"). The value is expressed in the following format: *access_key*/*YYYYMMDD*/*region*/*service*/aws4_request.

For more information, see Create a signed AWS API request in the *IAM User Guide*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-Date

The date that is used to create the signature. The format must be ISO 8601 basic format (YYYYMMDD'T'HHMMSS'Z'). For example, the following date time is a valid X-Amz-Date value: `20120325T120000Z`.

Condition: X-Amz-Date is optional for all requests; it can be used to override the date used for signing requests. If the Date header is specified in the ISO 8601 basic format, X-Amz-Date is not required. When X-Amz-Date is used, it always overrides the value of the Date header. For more information, see Elements of an AWS API request signature in the *IAM User Guide*.

Type: string

Required: Conditional

X-Amz-Security-Token

The temporary security token that was obtained through a call to AWS Security Token Service (AWS STS). For a list of services that support temporary security credentials from AWS STS, see AWS services that work with IAM in the *IAM User Guide*.

Condition: If you're using temporary security credentials from AWS STS, you must include the security token.

Type: string

Required: Conditional

X-Amz-Signature

Specifies the hex-encoded signature that was calculated from the string to sign and the derived signing key.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-SignedHeaders

Specifies all the HTTP headers that were included as part of the canonical request. For more information about specifying signed headers, see Create a signed AWS API request in the *IAM User Guide*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

Common Error Types

This section lists common error types that this AWS service may return. Not all services return all error types listed here. For errors specific to an API action for this service, see the topic for that API action.

AccessDeniedException

You don't have permission to perform this action. Verify that your IAM policy includes the required permissions.

HTTP Status Code: 403

ExpiredTokenException

The security token included in the request has expired. Request a new security token and try again.

HTTP Status Code: 403

IncompleteSignature

The request signature doesn't conform to AWS standards. Verify that you're using valid AWS credentials and that your request is properly formatted. If you're using an SDK, ensure it's up to date.

HTTP Status Code: 403

InternalFailure

The request can't be processed right now because of an internal server issue. Try again later. If the problem persists, contact AWS Support.

HTTP Status Code: 500

MalformedHttpRequestException

The request body can't be processed. This typically happens when the request body can't be decompressed using the specified content encoding algorithm. Verify that the content encoding header matches the compression format used.

HTTP Status Code: 400

NotAuthorized

You don't have permissions to perform this action. Verify that your IAM policy includes the required permissions.

HTTP Status Code: 401

OptInRequired

Your AWS account needs a subscription for this service. Verify that you've enabled the service in your account.

HTTP Status Code: 403

RequestAbortedException

The request was aborted before a response could be returned. This typically happens when the client closes the connection.

HTTP Status Code: 400

RequestEntityTooLargeException

The request entity is too large. Reduce the size of the request body and try again.

HTTP Status Code: 413

RequestTimeoutException

The request timed out. The server didn't receive the complete request within the expected time frame. Try again.

HTTP Status Code: 408

ServiceUnavailable

The service is temporarily unavailable. Try again later.

HTTP Status Code: 503

ThrottlingException

Your request rate is too high. The AWS SDKs automatically retry requests that receive this exception. Reduce the frequency of requests.

HTTP Status Code: 400

UnknownOperationException

The action or operation isn't recognized. Verify that the action name is spelled correctly and that it's supported by the API version you're using.

HTTP Status Code: 404

UnrecognizedClientException

The X.509 certificate or AWS access key ID you provided doesn't exist in our records. Verify that you're using valid credentials and that they haven't expired.

HTTP Status Code: 403

ValidationError

The input doesn't meet the required format or constraints. Check that all required parameters are included and that values are valid.

HTTP Status Code: 400