aws

API Reference

Application Auto Scaling

API Version 2016-02-06

Application Auto Scaling: API Reference

Table of Contents

Welcome

This is the *Application Auto Scaling API Reference*. With Application Auto Scaling, you can configure automatic scaling for the following resources:

- Amazon AppStream 2.0 fleets
- Amazon Aurora Replicas
- Amazon Comprehend document classification and entity recognizer endpoints
- Amazon DynamoDB tables and global secondary indexes throughput capacity
- Amazon ECS services
- Amazon ElastiCache replication groups (Redis OSS and Valkey) and Memcached clusters
- Amazon EMR clusters
- Amazon Keyspaces (for Apache Cassandra) tables
- AWS Lambda function provisioned concurrency
- Amazon Managed Streaming for Apache Kafka broker storage
- Amazon Neptune clusters
- Amazon SageMaker endpoint variants
- Amazon SageMaker inference components
- Amazon SageMaker serverless endpoint provisioned concurrency
- Spot Fleets (Amazon EC2)
- Pool of WorkSpaces
- Custom resources provided by your own applications or services

To learn more about Application Auto Scaling, see the Application Auto Scaling User Guide.

API Summary

The Application Auto Scaling service API includes three key sets of actions:

- Register and manage scalable targets - Register AWS or custom resources as scalable targets (a resource that Application Auto Scaling can scale), set minimum and maximum capacity limits, and retrieve information on existing scalable targets.

- Configure and manage automatic scaling - Define scaling policies to dynamically scale your resources in response to CloudWatch alarms, schedule one-time or recurring scaling actions, and retrieve your recent scaling activity history.
- Suspend and resume scaling - Temporarily suspend and later resume automatic scaling by calling the RegisterScalableTarget API action for any Application Auto Scaling scalable target. You can suspend and resume (individually or in combination) scale-out activities that are triggered by a scaling policy, scale-in activities that are triggered by a scaling policy, and scheduled scaling.

The documentation for each action shows the request syntax, the request parameters, and the response elements and provides links to language-specific SDK reference topics. For more information, see AWS SDKs.

API request rate

Application Auto Scaling uses the token bucket algorithm to implement API throttling. With this algorithm, your account has a bucket that holds a specific number of tokens. The number of tokens in the bucket represents your throttling limit at any given second. Application Auto Scaling throttles API requests based on a shared API bucket. For example, calls to the DescribeScalableTargets and DescribeScheduledActions API operations use tokens from the same bucket. Throttling means that Application Auto Scaling rejects a request because the request exceeds the service's limit for the number of requests per second. When a request is throttled, Application Auto Scaling returns a `RateExceeded` error. For more information, see My Auto Scaling API calls are getting throttled. What can I do to avoid this? in the AWS Knowledge Center.

This document was last published on April 10, 2026.

Actions

The following actions are supported:

- DeleteScalingPolicy
- DeleteScheduledAction
- DeregisterScalableTarget
- DescribeScalableTargets
- DescribeScalingActivities
- DescribeScalingPolicies
- DescribeScheduledActions
- GetPredictiveScalingForecast
- ListTagsForResource
- PutScalingPolicy
- PutScheduledAction
- RegisterScalableTarget
- TagResource
- UntagResource

DeleteScalingPolicy

Deletes the specified scaling policy for an Application Auto Scaling scalable target.

Deleting a step scaling policy deletes the underlying alarm action, but does not delete the CloudWatch alarm associated with the scaling policy, even if it no longer has an associated action.

For more information, see Delete a step scaling policy and Delete a target tracking scaling policy in the *Application Auto Scaling User Guide*.

Request Syntax

```
{
   "PolicyName": "string",
   "ResourceId": "string",
   "ScalableDimension": "string",
   "ServiceNamespace": "string"
}
```

Request Parameters

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

The request accepts the following data in JSON format.

PolicyName

The name of the scaling policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ResourceId

The identifier of the resource associated with the scalable target. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.

- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension

The scalable dimension. This string consists of the service namespace, resource type, and scaling property.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.

- dynamodb:table:WriteCapacityUnits - The provisioned write capacity for a DynamoDB table.
- dynamodb:index:ReadCapacityUnits - The provisioned read capacity for a DynamoDB global secondary index.
- dynamodb:index:WriteCapacityUnits - The provisioned write capacity for a DynamoDB global secondary index.
- rds:cluster:ReadReplicaCount - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- sagemaker:variant:DesiredInstanceCount - The number of EC2 instances for a SageMaker model endpoint variant.
- custom-resource:ResourceType:Property - The scalable dimension for a custom resource provided by your own application or service.
- comprehend:document-classifier-endpoint:DesiredInferenceUnits - The number of inference units for an Amazon Comprehend document classification endpoint.
- comprehend:entity-recognizer-endpoint:DesiredInferenceUnits - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- lambda:function:ProvisionedConcurrency - The provisioned concurrency for a Lambda function.
- cassandra:table:ReadCapacityUnits - The provisioned read capacity for an Amazon Keyspaces table.
- cassandra:table:WriteCapacityUnits - The provisioned write capacity for an Amazon Keyspaces table.
- kafka:broker-storage:VolumeSize - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- elasticache:cache-cluster:Nodes - The number of nodes for an Amazon ElastiCache cache cluster.
- elasticache:replication-group:NodeGroups - The number of node groups for an Amazon ElastiCache replication group.
- elasticache:replication-group:Replicas - The number of replicas per node group for an Amazon ElastiCache replication group.
- neptune:cluster:ReadReplicaCount - The count of read replicas in an Amazon Neptune DB cluster.

- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: Yes

ServiceNamespace

The namespace of the AWS service that provides the resource. For a resource provided by your own application or service, use `custom-resource` instead.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

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.

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to an Application Auto Scaling resource that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

ObjectNotFoundException

The specified object could not be found. For any operation that depends on the existence of a scalable target, this exception is thrown if the scalable target with the specified service namespace, resource ID, and scalable dimension does not exist. For any operation that deletes or deregisters a resource, this exception is thrown if the resource cannot be found.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example

The following example deletes a scaling policy for the Amazon ECS service `web-app` running in the `default` cluster.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.DeleteScalingPolicy
X-Amz-Date: 20190506T205712Z
User-Agent: aws-cli/1.10.23 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "PolicyName": "my-scale-out-policy",
    "ServiceNamespace": "ecs",
    "ScalableDimension": "ecs:service:DesiredCount",
    "ResourceId": "service/my-cluster/my-service"
}
```

See Also

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

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

- AWS SDK for Ruby V3

DeleteScheduledAction

Deletes the specified scheduled action for an Application Auto Scaling scalable target.

For more information, see Delete a scheduled action in the *Application Auto Scaling User Guide*.

Request Syntax

```
{
   "ResourceId": "string",
   "ScalableDimension": "string",
   "ScheduledActionName": "string",
   "ServiceNamespace": "string"
}
```

Request Parameters

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

The request accepts the following data in JSON format.

ResourceId

The identifier of the resource associated with the scheduled action. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.

- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.

- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension

The scalable dimension. This string consists of the service namespace, resource type, and scaling property.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- `sagemaker:variant:DesiredInstanceCount` - The number of EC2 instances for a SageMaker model endpoint variant.
- `custom-resource:ResourceType:Property` - The scalable dimension for a custom resource provided by your own application or service.

- `comprehend:document-classifier-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend document classification endpoint.
- `comprehend:entity-recognizer-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- `lambda:function:ProvisionedConcurrency` - The provisioned concurrency for a Lambda function.
- `cassandra:table:ReadCapacityUnits` - The provisioned read capacity for an Amazon Keyspaces table.
- `cassandra:table:WriteCapacityUnits` - The provisioned write capacity for an Amazon Keyspaces table.
- `kafka:broker-storage:VolumeSize` - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- `elasticache:cache-cluster:Nodes` - The number of nodes for an Amazon ElastiCache cache cluster.
- `elasticache:replication-group:NodeGroups` - The number of node groups for an Amazon ElastiCache replication group.
- `elasticache:replication-group:Replicas` - The number of replicas per node group for an Amazon ElastiCache replication group.
- `neptune:cluster:ReadReplicaCount` - The count of read replicas in an Amazon Neptune DB cluster.
- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-`

`resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: Yes

ScheduledActionName

The name of the scheduled action.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ServiceNamespace

The namespace of the AWS service that provides the resource. For a resource provided by your own application or service, use `custom-resource` instead.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

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.

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to an Application Auto Scaling resource that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

ObjectNotFoundException

The specified object could not be found. For any operation that depends on the existence of a scalable target, this exception is thrown if the scalable target with the specified service namespace, resource ID, and scalable dimension does not exist. For any operation that deletes or deregisters a resource, this exception is thrown if the resource cannot be found.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example

The following example deletes a scheduled action for the AppStream 2.0 fleet called `sample-fleet`.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.DeleteScheduledAction
X-Amz-Date: 20190506T205712Z
User-Agent: aws-cli/1.10.23 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "ScheduledActionName": "my-recurring-action",
    "ServiceNamespace": "appstream",
    "ScalableDimension": "appstream:fleet:DesiredCapacity",
    "ResourceId": "fleet/sample-fleet"
}
```

See Also

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

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

DeregisterScalableTarget

Deregisters an Application Auto Scaling scalable target when you have finished using it. To see which resources have been registered, use DescribeScalableTargets.

> **Note**
>
> Deregistering a scalable target deletes the scaling policies and the scheduled actions that are associated with it.

Request Syntax

```
{
   "ResourceId": "string",
   "ScalableDimension": "string",
   "ServiceNamespace": "string"
}
```

Request Parameters

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

The request accepts the following data in JSON format.

ResourceId

The identifier of the resource associated with the scalable target. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.

- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.

- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension

The scalable dimension associated with the scalable target. This string consists of the service namespace, resource type, and scaling property.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- `sagemaker:variant:DesiredInstanceCount` - The number of EC2 instances for a SageMaker model endpoint variant.

- `custom-resource:ResourceType:Property` - The scalable dimension for a custom resource provided by your own application or service.
- `comprehend:document-classifier-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend document classification endpoint.
- `comprehend:entity-recognizer-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- `lambda:function:ProvisionedConcurrency` - The provisioned concurrency for a Lambda function.
- `cassandra:table:ReadCapacityUnits` - The provisioned read capacity for an Amazon Keyspaces table.
- `cassandra:table:WriteCapacityUnits` - The provisioned write capacity for an Amazon Keyspaces table.
- `kafka:broker-storage:VolumeSize` - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- `elasticache:cache-cluster:Nodes` - The number of nodes for an Amazon ElastiCache cache cluster.
- `elasticache:replication-group:NodeGroups` - The number of node groups for an Amazon ElastiCache replication group.
- `elasticache:replication-group:Replicas` - The number of replicas per node group for an Amazon ElastiCache replication group.
- `neptune:cluster:ReadReplicaCount` - The count of read replicas in an Amazon Neptune DB cluster.
- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits`

`| dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: Yes

ServiceNamespace

The namespace of the AWS service that provides the resource. For a resource provided by your own application or service, use `custom-resource` instead.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

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.

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to an Application Auto Scaling resource that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

ObjectNotFoundException

The specified object could not be found. For any operation that depends on the existence of a scalable target, this exception is thrown if the scalable target with the specified service namespace, resource ID, and scalable dimension does not exist. For any operation that deletes or deregisters a resource, this exception is thrown if the resource cannot be found.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example

The following example deregisters a scalable target for an Amazon ECS service called `web-app` that is running in the `default` cluster.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.DeregisterScalableTarget
```

```
X-Amz-Date: 20190506T210150Z
User-Agent: aws-cli/1.10.23 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "ResourceId": "service/my-cluster/my-service",
    "ServiceNamespace": "ecs",
    "ScalableDimension": "ecs:service:DesiredCount"
}
```

See Also

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

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

DescribeScalableTargets

Gets information about the scalable targets in the specified namespace.

You can filter the results using `ResourceIds` and `ScalableDimension`.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string",
   "ResourceIds": [ "string" ],
   "ScalableDimension": "string",
   "ServiceNamespace": "string"
}
```

Request 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 scalable targets. This value can be between 1 and 50. The default value is 50.

If this parameter is used, the operation returns up to `MaxResults` results at a time, along with a `NextToken` value. To get the next set of results, include the `NextToken` value in a subsequent call. If this parameter is not used, the operation returns up to 50 results and a `NextToken` value, if applicable.

Type: Integer

Required: No

NextToken

The token for the next set of results.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

ResourceIds

The identifier of the resource associated with the scalable target. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.

- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: Array of strings

Array Members: Maximum number of 50 items.

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

ScalableDimension

The scalable dimension associated with the scalable target. This string consists of the service namespace, resource type, and scaling property. If you specify a scalable dimension, you must also specify a resource ID.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- `sagemaker:variant:DesiredInstanceCount` - The number of EC2 instances for a SageMaker model endpoint variant.
- `custom-resource:ResourceType:Property` - The scalable dimension for a custom resource provided by your own application or service.
- `comprehend:document-classifier-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend document classification endpoint.
- `comprehend:entity-recognizer-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- `lambda:function:ProvisionedConcurrency` - The provisioned concurrency for a Lambda function.
- `cassandra:table:ReadCapacityUnits` - The provisioned read capacity for an Amazon Keyspaces table.
- `cassandra:table:WriteCapacityUnits` - The provisioned write capacity for an Amazon Keyspaces table.
- `kafka:broker-storage:VolumeSize` - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- `elasticache:cache-cluster:Nodes` - The number of nodes for an Amazon ElastiCache cache cluster.

- `elasticache:replication-group:NodeGroups` - The number of node groups for an Amazon ElastiCache replication group.
- `elasticache:replication-group:Replicas` - The number of replicas per node group for an Amazon ElastiCache replication group.
- `neptune:cluster:ReadReplicaCount` - The count of read replicas in an Amazon Neptune DB cluster.
- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: No

ServiceNamespace

The namespace of the AWS service that provides the resource. For a resource provided by your own application or service, use `custom-resource` instead.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

Response Syntax

```
{
   "NextToken": "string",
   "ScalableTargets": [
      {
         "CreationTime": number,
         "MaxCapacity": number,
         "MinCapacity": number,
         "PredictedCapacity": number,
         "ResourceId": "string",
         "RoleARN": "string",
         "ScalableDimension": "string",
         "ScalableTargetARN": "string",
         "ServiceNamespace": "string",
         "SuspendedState": {
            "DynamicScalingInSuspended": boolean,
            "DynamicScalingOutSuspended": boolean,
            "ScheduledScalingSuspended": 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.

NextToken

The token required to get the next set of results. This value is `null` if there are no more results to return.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

ScalableTargets

The scalable targets that match the request parameters.

Type: Array of ScalableTarget objects

Errors

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

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to an Application Auto Scaling resource that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

InvalidNextTokenException

The next token supplied was invalid.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in

the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example

The following example describes the scalable targets for the `ecs` service namespace.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.DescribeScalableTargets
X-Amz-Date: 20190506T184921Z
User-Agent: aws-cli/1.10.23 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "ServiceNamespace": "ecs"
}
```

Sample Response

```
HTTP/1.1 200 OK
x-amzn-RequestId: [request-id]
Content-Type: application/x-amz-json-1.1
Content-Length: 272
Date: Fri, 06 May 2019 18:49:21 GMT

{
    "ScalableTargets": [
        {
            "CreationTime": 1462558906.199,
            "MaxCapacity": 10,
            "MinCapacity": 1,
            "ResourceId": "service/my-cluster/my-service",
            "RoleARN": "arn:aws:iam::012345678910:role/
aws-service-role/ecs.application-autoscaling.amazonaws.com/
AWSServiceRoleForApplicationAutoScaling_ECSService",
            "ScalableDimension": "ecs:service:DesiredCount",
            "ServiceNamespace": "ecs",
```

```
            "SuspendedState": {
                "DynamicScalingInSuspended": false,
                "DynamicScalingOutSuspended": false,
                "ScheduledScalingSuspended": false
            }
        }
    ]
}
```

See Also

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

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

DescribeScalingActivities

Provides descriptive information about the scaling activities in the specified namespace from the previous six weeks.

You can filter the results using `ResourceId` and `ScalableDimension`.

For information about viewing scaling activities using the AWS CLI, see Scaling activities for Application Auto Scaling.

Request Syntax

```
{
   "IncludeNotScaledActivities": boolean,
   "MaxResults": number,
   "NextToken": "string",
   "ResourceId": "string",
   "ScalableDimension": "string",
   "ServiceNamespace": "string"
}
```

Request Parameters

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

The request accepts the following data in JSON format.

IncludeNotScaledActivities

Specifies whether to include activities that aren't scaled (*not scaled activities*) in the response. Not scaled activities are activities that aren't completed or started for various reasons, such as preventing infinite scaling loops. For help interpreting the not scaled reason details in the response, see Scaling activities for Application Auto Scaling.

Type: Boolean

Required: No

MaxResults

The maximum number of scalable targets. This value can be between 1 and 50. The default value is 50.

If this parameter is used, the operation returns up to `MaxResults` results at a time, along with a `NextToken` value. To get the next set of results, include the `NextToken` value in a subsequent call. If this parameter is not used, the operation returns up to 50 results and a `NextToken` value, if applicable.

Type: Integer

Required: No

NextToken

The token for the next set of results.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

ResourceId

The identifier of the resource associated with the scaling activity. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.

- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

ScalableDimension

The scalable dimension. This string consists of the service namespace, resource type, and scaling property. If you specify a scalable dimension, you must also specify a resource ID.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- `sagemaker:variant:DesiredInstanceCount` - The number of EC2 instances for a SageMaker model endpoint variant.
- `custom-resource:ResourceType:Property` - The scalable dimension for a custom resource provided by your own application or service.
- `comprehend:document-classifier-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend document classification endpoint.
- `comprehend:entity-recognizer-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend entity recognizer endpoint.

- `lambda:function:ProvisionedConcurrency` - The provisioned concurrency for a Lambda function.
- `cassandra:table:ReadCapacityUnits` - The provisioned read capacity for an Amazon Keyspaces table.
- `cassandra:table:WriteCapacityUnits` - The provisioned write capacity for an Amazon Keyspaces table.
- `kafka:broker-storage:VolumeSize` - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- `elasticache:cache-cluster:Nodes` - The number of nodes for an Amazon ElastiCache cache cluster.
- `elasticache:replication-group:NodeGroups` - The number of node groups for an Amazon ElastiCache replication group.
- `elasticache:replication-group:Replicas` - The number of replicas per node group for an Amazon ElastiCache replication group.
- `neptune:cluster:ReadReplicaCount` - The count of read replicas in an Amazon Neptune DB cluster.
- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits`

| kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions

Required: No

ServiceNamespace

The namespace of the AWS service that provides the resource. For a resource provided by your own application or service, use `custom-resource` instead.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

Response Syntax

```
{
   "NextToken": "string",
   "ScalingActivities": [
      {
         "ActivityId": "string",
         "Cause": "string",
         "Description": "string",
         "Details": "string",
         "EndTime": number,
         "NotScaledReasons": [
            {
               "Code": "string",
               "CurrentCapacity": number,
               "MaxCapacity": number,
               "MinCapacity": number
            }
         ],
         "ResourceId": "string",
```

```
            "ScalableDimension": "string",
            "ServiceNamespace": "string",
            "StartTime": number,
            "StatusCode": "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.

NextToken

The token required to get the next set of results. This value is `null` if there are no more results to return.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

ScalingActivities

A list of scaling activity objects.

Type: Array of ScalingActivity objects

Errors

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

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to an Application Auto Scaling resource that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

InvalidNextTokenException

The next token supplied was invalid.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example of scaling activities for a scaling policy

The following example describes the scaling activities for an Amazon ECS service named `web-app` that's running in the `default` cluster. It shows the scaling activities for the scaling policy named `cpu75-target-tracking-scaling-policy`, initiated by the CloudWatch alarm named `TargetTracking-service/my-cluster/my-service-AlarmHigh-d4f0770c-b46e-434a-a60f-3b36d653feca`.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.DescribeScalingActivities
X-Amz-Date: 20190506T224112Z
User-Agent: aws-cli/1.10.26 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS
```

```
{
  "ResourceId": "service/my-cluster/my-service",
  "ServiceNamespace": "ecs",
  "ScalableDimension": "ecs:service:DesiredCount"
}
```

Sample Response

```
HTTP/1.1 200 OK
x-amzn-RequestId: [request-id]
Content-Type: application/x-amz-json-1.1
Content-Length: 1784
Date: Fri, 06 May 2019 22:41:12 GMT

{
  "ScalingActivities": [
    {
      "ScalableDimension": "ecs:service:DesiredCount",
      "Description": "Setting desired count to 3.",
      "ResourceId": "service/my-cluster/my-service",
      "ActivityId": "4d759079-a31f-4d0c-8468-504c56e2eecf",
      "StartTime": 1462574194.658,
      "ServiceNamespace": "ecs",
      "EndTime": 1462574276.686,
      "Cause": "monitor alarm TargetTracking-service/my-cluster/my-service-AlarmHigh-
d4f0770c-b46e-434a-a60f-3b36d653feca in state ALARM triggered policy cpu75-target-
tracking-scaling-policy",
      "StatusMessage": "Successfully set desired count to 3. Change successfully
 fulfilled by ecs.",
      "StatusCode": "Successful"
    },
    {
      "ScalableDimension": "ecs:service:DesiredCount",
      "Description": "Setting desired count to 2.",
      "ResourceId": "service/my-cluster/my-service",
      "ActivityId": "90aff0eb-dd6a-443c-889b-b809e78061c1",
      "StartTime": 1462574254.223,
      "ServiceNamespace": "ecs",
      "EndTime": 1462574333.492,
      "Cause": "monitor alarm TargetTracking-service/my-cluster/my-service-AlarmHigh-
d4f0770c-b46e-434a-a60f-3b36d653feca in state ALARM triggered policy cpu75-target-
tracking-scaling-policy",
```

```
      "StatusMessage": "Successfully set desired count to 2. Change successfully
 fulfilled by ecs.",
      "StatusCode": "Successful"
    }
  ]
}
```

Example of scaling activities for scheduled actions

The following example describes the scaling activities for a DynamoDB table named `my-table`. It shows the scaling activities for scheduled actions named `my-first-scheduled-action` and `my-second-scheduled-action`.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.DescribeScalingActivities
X-Amz-Date: 20190526T110828Z
User-Agent: aws-cli/1.10.26 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
  "ResourceId": "table/my-table",
  "ServiceNamespace": "dynamodb",
  "ScalableDimension": "dynamodb:table:WriteCapacityUnits"
}
```

Sample Response

```
HTTP/1.1 200 OK
x-amzn-RequestId: [request-id]
Content-Type: application/x-amz-json-1.1
Content-Length: 1784
Date: Fri, 26 May 2019 11:08:28 GMT

{
  "ScalingActivities": [
    {
```

```
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Description": "Setting write capacity units to 10.",
      "ResourceId": "table/my-table",
      "ActivityId": "4d1308c0-bbcf-4514-a673-b0220ae38547",
      "StartTime": 1561574415.086,
      "ServiceNamespace": "dynamodb",
      "Cause": "maximum capacity was set to 10",
      "StatusMessage": "Successfully set write capacity units to 10. Waiting for change
 to be fulfilled by dynamodb.",
      "StatusCode": "InProgress"
    },
    {
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Description": "Setting min capacity to 5 and max capacity to 10",
      "ResourceId": "table/my-table",
      "ActivityId": "f2b7847b-721d-4e01-8ef0-0c8d3bacc1c7",
      "StartTime": 1561574414.644,
      "ServiceNamespace": "dynamodb",
      "Cause": "scheduled action name my-second-scheduled-action was triggered",
      "StatusMessage": "Successfully set min capacity to 5 and max capacity to 10",
      "StatusCode": "Successful"
    },
    {
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Description": "Setting write capacity units to 15.",
      "ResourceId": "table/my-table",
      "ActivityId": "d8ea4de6-9eaa-499f-b466-2cc5e681ba8b",
      "StartTime": 1561574108.904,
      "ServiceNamespace": "dynamodb",
      "EndTime": 1561574140.255,
      "Cause": "minimum capacity was set to 15",
      "StatusMessage": "Successfully set write capacity units to 15. Change
 successfully fulfilled by dynamodb.",
      "StatusCode": "Successful"
    },
    {
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Description": "Setting min capacity to 15 and max capacity to 20",
      "ResourceId": "table/my-table",
      "ActivityId": "3250fd06-6940-4e8e-bb1f-d494db7554d2",
      "StartTime": 1561574108.512,
      "ServiceNamespace": "dynamodb",
      "Cause": "scheduled action name my-first-scheduled-action was triggered",
      "StatusMessage": "Successfully set min capacity to 15 and max capacity to 20",
```

```
      "StatusCode": "Successful"
    }
  ]
}
```

See Also

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

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

DescribeScalingPolicies

Describes the Application Auto Scaling scaling policies for the specified service namespace.

You can filter the results using `ResourceId`, `ScalableDimension`, and `PolicyNames`.

For more information, see Target tracking scaling policies and Step scaling policies in the *Application Auto Scaling User Guide*.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string",
   "PolicyNames": [ "string" ],
   "ResourceId": "string",
   "ScalableDimension": "string",
   "ServiceNamespace": "string"
}
```

Request 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 scalable targets. This value can be between 1 and 10. The default value is 10.

If this parameter is used, the operation returns up to `MaxResults` results at a time, along with a `NextToken` value. To get the next set of results, include the `NextToken` value in a subsequent call. If this parameter is not used, the operation returns up to 10 results and a `NextToken` value, if applicable.

Type: Integer

Required: No

NextToken

The token for the next set of results.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

PolicyNames

The names of the scaling policies to describe.

Type: Array of strings

Array Members: Maximum number of 50 items.

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

ResourceId

The identifier of the resource associated with the scaling policy. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.

- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

ScalableDimension

The scalable dimension. This string consists of the service namespace, resource type, and scaling property. If you specify a scalable dimension, you must also specify a resource ID.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- `sagemaker:variant:DesiredInstanceCount` - The number of EC2 instances for a SageMaker model endpoint variant.
- `custom-resource:ResourceType:Property` - The scalable dimension for a custom resource provided by your own application or service.
- `comprehend:document-classifier-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend document classification endpoint.
- `comprehend:entity-recognizer-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend entity recognizer endpoint.

- `lambda:function:ProvisionedConcurrency` - The provisioned concurrency for a Lambda function.
- `cassandra:table:ReadCapacityUnits` - The provisioned read capacity for an Amazon Keyspaces table.
- `cassandra:table:WriteCapacityUnits` - The provisioned write capacity for an Amazon Keyspaces table.
- `kafka:broker-storage:VolumeSize` - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- `elasticache:cache-cluster:Nodes` - The number of nodes for an Amazon ElastiCache cache cluster.
- `elasticache:replication-group:NodeGroups` - The number of node groups for an Amazon ElastiCache replication group.
- `elasticache:replication-group:Replicas` - The number of replicas per node group for an Amazon ElastiCache replication group.
- `neptune:cluster:ReadReplicaCount` - The count of read replicas in an Amazon Neptune DB cluster.
- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits`

| kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions

Required: No

ServiceNamespace

The namespace of the AWS service that provides the resource. For a resource provided by your own application or service, use `custom-resource` instead.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

Response Syntax

```
{
   "NextToken": "string",
   "ScalingPolicies": [
      {
         "Alarms": [
            {
               "AlarmARN": "string",
               "AlarmName": "string"
            }
         ],
         "CreationTime": number,
         "PolicyARN": "string",
         "PolicyName": "string",
         "PolicyType": "string",
         "PredictiveScalingPolicyConfiguration": {
            "MaxCapacityBreachBehavior": "string",
            "MaxCapacityBuffer": number,
            "MetricSpecifications": [
```

```
{
   "CustomizedCapacityMetricSpecification": {
      "MetricDataQueries": [
         {
            "Expression": "string",
            "Id": "string",
            "Label": "string",
            "MetricStat": {
               "Metric": {
                  "Dimensions": [
                     {
                        "Name": "string",
                        "Value": "string"
                     }
                  ],
                  "MetricName": "string",
                  "Namespace": "string"
               },
               "Stat": "string",
               "Unit": "string"
            },
            "ReturnData": boolean
         }
      ]
   },
   "CustomizedLoadMetricSpecification": {
      "MetricDataQueries": [
         {
            "Expression": "string",
            "Id": "string",
            "Label": "string",
            "MetricStat": {
               "Metric": {
                  "Dimensions": [
                     {
                        "Name": "string",
                        "Value": "string"
                     }
                  ],
                  "MetricName": "string",
                  "Namespace": "string"
               },
               "Stat": "string",
               "Unit": "string"
```

```
                        },
                        "ReturnData": boolean
                     }
                  ]
               },
               "CustomizedScalingMetricSpecification": {
                  "MetricDataQueries": [
                     {
                        "Expression": "string",
                        "Id": "string",
                        "Label": "string",
                        "MetricStat": {
                           "Metric": {
                              "Dimensions": [
                                 {
                                    "Name": "string",
                                    "Value": "string"
                                 }
                              ],
                              "MetricName": "string",
                              "Namespace": "string"
                           },
                           "Stat": "string",
                           "Unit": "string"
                        },
                        "ReturnData": boolean
                     }
                  ]
               },
               "PredefinedLoadMetricSpecification": {
                  "PredefinedMetricType": "string",
                  "ResourceLabel": "string"
               },
               "PredefinedMetricPairSpecification": {
                  "PredefinedMetricType": "string",
                  "ResourceLabel": "string"
               },
               "PredefinedScalingMetricSpecification": {
                  "PredefinedMetricType": "string",
                  "ResourceLabel": "string"
               },
               "TargetValue": number
            }
         ],
```

```
            "Mode": "string",
            "SchedulingBufferTime": number
         },
         "ResourceId": "string",
         "ScalableDimension": "string",
         "ServiceNamespace": "string",
         "StepScalingPolicyConfiguration": {
            "AdjustmentType": "string",
            "Cooldown": number,
            "MetricAggregationType": "string",
            "MinAdjustmentMagnitude": number,
            "StepAdjustments": [
               {
                  "MetricIntervalLowerBound": number,
                  "MetricIntervalUpperBound": number,
                  "ScalingAdjustment": number
               }
            ]
         },
         "TargetTrackingScalingPolicyConfiguration": {
            "CustomizedMetricSpecification": {
               "Dimensions": [
                  {
                     "Name": "string",
                     "Value": "string"
                  }
               ],
               "MetricName": "string",
               "Metrics": [
                  {
                     "Expression": "string",
                     "Id": "string",
                     "Label": "string",
                     "MetricStat": {
                        "Metric": {
                           "Dimensions": [
                              {
                                 "Name": "string",
                                 "Value": "string"
                              }
                           ],
                           "MetricName": "string",
                           "Namespace": "string"
                        },
```

```
                        "Stat": "string",
                        "Unit": "string"
                     },
                     "ReturnData": boolean
                  }
               ],
               "Namespace": "string",
               "Statistic": "string",
               "Unit": "string"
            },
            "DisableScaleIn": boolean,
            "PredefinedMetricSpecification": {
               "PredefinedMetricType": "string",
               "ResourceLabel": "string"
            },
            "ScaleInCooldown": number,
            "ScaleOutCooldown": number,
            "TargetValue": 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

The token required to get the next set of results. This value is `null` if there are no more results to return.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

ScalingPolicies

Information about the scaling policies.

Type: Array of ScalingPolicy objects

Errors

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

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to an Application Auto Scaling resource that already has a pending update.

HTTP Status Code: 400

FailedResourceAccessException

Failed access to resources caused an exception. This exception is thrown when Application Auto Scaling is unable to retrieve the alarms associated with a scaling policy due to a client error, for example, if the role ARN specified for a scalable target does not have permission to call the CloudWatch DescribeAlarms on your behalf.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

InvalidNextTokenException

The next token supplied was invalid.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example

The following example describes the scaling policies for the ecs service namespace.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.DescribeScalingPolicies
X-Amz-Date: 20190506T194435Z
User-Agent: aws-cli/1.10.23 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "ServiceNamespace": "ecs"
}
```

Sample Response

```
HTTP/1.1 200 OK
x-amzn-RequestId: [request-id]
Content-Type: application/x-amz-json-1.1
Content-Length: 1393
Date: Fri, 06 May 2019 19:44:35 GMT

{
    "ScalingPolicies": [
        {
            "Alarms": [
                {
                    "AlarmARN": "arn:aws:cloudwatch:us-west-2:012345678910:alarm:step-
scaling-alarmhigh-ecs:service/my-cluster/my-service",
                    "AlarmName": "Step-Scaling-AlarmHigh-ECS:service/my-cluster/my-
service"
                }
            ],
            "CreationTime": 1462561899.23,
            "PolicyARN": "arn:aws:autoscaling:us-
west-2:012345678910:scalingPolicy:ac542982-cbeb-4294-891c-a5a941dfa787:resource/ecs/
service/my-cluster/my-service:policyName/my-scale-out-policy",
```

```
            "PolicyName": "my-scale-out-policy",
            "PolicyType": "StepScaling",
            "ResourceId": "service/my-cluster/my-service",
            "ScalableDimension": "ecs:service:DesiredCount",
            "ServiceNamespace": "ecs",
            "StepScalingPolicyConfiguration": {
                "AdjustmentType": "PercentChangeInCapacity",
                "Cooldown": 60,
                "MetricAggregationType": "Average",
                "StepAdjustments": [
                    {
                        "MetricIntervalLowerBound": 0,
                        "ScalingAdjustment": 200
                    }
                ]
            }
        },
        {
            "Alarms": [
                {
                    "AlarmARN": "arn:aws:cloudwatch:us-west-2:012345678910:alarm:step-
scaling-alarmlow-ecs:service/my-cluster/my-service",
                    "AlarmName": "Step-Scaling-AlarmLow-ECS:service/my-cluster/my-
service"
                }
            ],
            "CreationTime": 1462562575.099,
            "PolicyARN": "arn:aws:autoscaling:us-
west-2:012345678910:scalingPolicy:6d8972f3-efc8-437c-92d1-6270f29a66e7:resource/ecs/
service/my-cluster/my-service:policyName/my-scale-in-policy",
            "PolicyName": "my-scale-in-policy",
            "PolicyType": "StepScaling",
            "ResourceId": "service/my-cluster/my-service",
            "ScalableDimension": "ecs:service:DesiredCount",
            "ServiceNamespace": "ecs",
            "StepScalingPolicyConfiguration": {
                "AdjustmentType": "PercentChangeInCapacity",
                "Cooldown": 120,
                "MetricAggregationType": "Average",
                "MinAdjustmentMagnitude": 1,
                "StepAdjustments": [
                    {
                        "MetricIntervalLowerBound": -15,
                        "MetricIntervalUpperBound": 0
```

```
                        "ScalingAdjustment": -25,
                    },
                    {
                        "MetricIntervalUpperBound": -15,
                        "ScalingAdjustment": -50
                    }
                ]
            }
        }
    ]
}
```

See Also

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

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

DescribeScheduledActions

Describes the Application Auto Scaling scheduled actions for the specified service namespace.

You can filter the results using the `ResourceId`, `ScalableDimension`, and `ScheduledActionNames` parameters.

For more information, see Scheduled scaling in the *Application Auto Scaling User Guide*.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string",
   "ResourceId": "string",
   "ScalableDimension": "string",
   "ScheduledActionNames": [ "string" ],
   "ServiceNamespace": "string"
}
```

Request 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 scheduled action results. This value can be between 1 and 50. The default value is 50.

If this parameter is used, the operation returns up to `MaxResults` results at a time, along with a `NextToken` value. To get the next set of results, include the `NextToken` value in a subsequent call. If this parameter is not used, the operation returns up to 50 results and a `NextToken` value, if applicable.

Type: Integer

Required: No

NextToken

The token for the next set of results.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

ResourceId

The identifier of the resource associated with the scheduled action. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.

- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

ScalableDimension

The scalable dimension. This string consists of the service namespace, resource type, and scaling property. If you specify a scalable dimension, you must also specify a resource ID.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- `sagemaker:variant:DesiredInstanceCount` - The number of EC2 instances for a SageMaker model endpoint variant.
- `custom-resource:ResourceType:Property` - The scalable dimension for a custom resource provided by your own application or service.
- `comprehend:document-classifier-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend document classification endpoint.
- `comprehend:entity-recognizer-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- `lambda:function:ProvisionedConcurrency` - The provisioned concurrency for a Lambda function.
- `cassandra:table:ReadCapacityUnits` - The provisioned read capacity for an Amazon Keyspaces table.
- `cassandra:table:WriteCapacityUnits` - The provisioned write capacity for an Amazon Keyspaces table.
- `kafka:broker-storage:VolumeSize` - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- `elasticache:cache-cluster:Nodes` - The number of nodes for an Amazon ElastiCache cache cluster.

- `elasticache:replication-group:NodeGroups` - The number of node groups for an Amazon ElastiCache replication group.
- `elasticache:replication-group:Replicas` - The number of replicas per node group for an Amazon ElastiCache replication group.
- `neptune:cluster:ReadReplicaCount` - The count of read replicas in an Amazon Neptune DB cluster.
- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: No

ScheduledActionNames

The names of the scheduled actions to describe.

Type: Array of strings

Array Members: Maximum number of 50 items.

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

ServiceNamespace

The namespace of the AWS service that provides the resource. For a resource provided by your own application or service, use `custom-resource` instead.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

Response Syntax

```
{
   "NextToken": "string",
   "ScheduledActions": [
      {
         "CreationTime": number,
         "EndTime": number,
         "ResourceId": "string",
         "ScalableDimension": "string",
         "ScalableTargetAction": {
            "MaxCapacity": number,
            "MinCapacity": number
         },
         "Schedule": "string",
         "ScheduledActionARN": "string",
         "ScheduledActionName": "string",
         "ServiceNamespace": "string",
         "StartTime": number,
         "Timezone": "string"
```

```
        }
    ]
}
```

Response Elements

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 required to get the next set of results. This value is `null` if there are no more results to return.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

ScheduledActions

Information about the scheduled actions.

Type: Array of ScheduledAction objects

Errors

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

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to an Application Auto Scaling resource that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

InvalidNextTokenException

The next token supplied was invalid.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example

The following example describes the scheduled actions for the dynamodb service namespace.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.DescribeScheduledActions
X-Amz-Date: 20190506T194435Z
User-Agent: aws-cli/1.10.23 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "ServiceNamespace": "dynamodb"
}
```

Sample Response

```
HTTP/1.1 200 OK
x-amzn-RequestId: [request-id]
Content-Type: application/x-amz-json-1.1
Content-Length: 834
```

```
Date: Fri, 06 May 2019 19:44:35 GMT

{
    "ScheduledActions": [
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Schedule": "at(2019-05-20T18:35:00)",
            "ResourceId": "table/my-table",
            "CreationTime": 1561571888.361,
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:2d36aa3b-cdf9-4565-b290-81db519b227d:resource/
dynamodb/table/my-table:scheduledActionName/my-first-scheduled-action",
            "ScalableTargetAction": {
                "MinCapacity": 15,
                "MaxCapacity": 20
            },
            "ScheduledActionName": "my-first-scheduled-action",
            "ServiceNamespace": "dynamodb"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Schedule": "at(2019-05-20T18:40:00)",
            "ResourceId": "table/my-table",
            "CreationTime": 1561571946.021,
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:2d36aa3b-cdf9-4565-b290-81db519b227d:resource/
dynamodb/table/my-table:scheduledActionName/my-second-scheduled-action",
            "ScalableTargetAction": {
                "MinCapacity": 5,
                "MaxCapacity": 10
            },
            "ScheduledActionName": "my-second-scheduled-action",
            "ServiceNamespace": "dynamodb"
        }
    ]
}
```

See Also

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

- AWS Command Line Interface V2

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

GetPredictiveScalingForecast

Retrieves the forecast data for a predictive scaling policy.

Load forecasts are predictions of the hourly load values using historical load data from CloudWatch and an analysis of historical trends. Capacity forecasts are represented as predicted values for the minimum capacity that is needed on an hourly basis, based on the hourly load forecast.

A minimum of 24 hours of data is required to create the initial forecasts. However, having a full 14 days of historical data results in more accurate forecasts.

Request Syntax

```
{
   "EndTime": number,
   "PolicyName": "string",
   "ResourceId": "string",
   "ScalableDimension": "string",
   "ServiceNamespace": "string",
   "StartTime": 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.

EndTime

The exclusive end time of the time range for the forecast data to get. The maximum time duration between the start and end time is 30 days.

Type: Timestamp

Required: Yes

PolicyName

The name of the policy.

Type: String

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

Pattern: `\p{Print}+`

Required: Yes

ResourceId

The identifier of the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension

The scalable dimension.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: Yes

ServiceNamespace

The namespace of the AWS service that provides the resource. For a resource provided by your own application or service, use `custom-resource` instead.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

StartTime

The inclusive start time of the time range for the forecast data to get. At most, the date and time can be one year before the current date and time

Type: Timestamp

Required: Yes

Response Syntax

```
{
   "CapacityForecast": {
      "Timestamps": [ number ],
      "Values": [ number ]
   },
   "LoadForecast": [
      {
         "MetricSpecification": {
            "CustomizedCapacityMetricSpecification": {
               "MetricDataQueries": [
                  {
                     "Expression": "string",
                     "Id": "string",
                     "Label": "string",
                     "MetricStat": {
                        "Metric": {
                           "Dimensions": [
                              {
                                 "Name": "string",
```

```
                        "Value": "string"
                     }
                  ],
                  "MetricName": "string",
                  "Namespace": "string"
               },
               "Stat": "string",
               "Unit": "string"
            },
            "ReturnData": boolean
         }
      ]
   },
   "CustomizedLoadMetricSpecification": {
      "MetricDataQueries": [
         {
            "Expression": "string",
            "Id": "string",
            "Label": "string",
            "MetricStat": {
               "Metric": {
                  "Dimensions": [
                     {
                        "Name": "string",
                        "Value": "string"
                     }
                  ],
                  "MetricName": "string",
                  "Namespace": "string"
               },
               "Stat": "string",
               "Unit": "string"
            },
            "ReturnData": boolean
         }
      ]
   },
   "CustomizedScalingMetricSpecification": {
      "MetricDataQueries": [
         {
            "Expression": "string",
            "Id": "string",
            "Label": "string",
            "MetricStat": {
```

```
                        "Metric": {
                           "Dimensions": [
                              {
                                 "Name": "string",
                                 "Value": "string"
                              }
                           ],
                           "MetricName": "string",
                           "Namespace": "string"
                        },
                        "Stat": "string",
                        "Unit": "string"
                     },
                     "ReturnData": boolean
                  }
               ]
            },
            "PredefinedLoadMetricSpecification": {
               "PredefinedMetricType": "string",
               "ResourceLabel": "string"
            },
            "PredefinedMetricPairSpecification": {
               "PredefinedMetricType": "string",
               "ResourceLabel": "string"
            },
            "PredefinedScalingMetricSpecification": {
               "PredefinedMetricType": "string",
               "ResourceLabel": "string"
            },
            "TargetValue": number
         },
         "Timestamps": [ number ],
         "Values": [ number ]
      }
   ],
   "UpdateTime": 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.

CapacityForecast

The capacity forecast.

Type: CapacityForecast object

LoadForecast

The load forecast.

Type: Array of LoadForecast objects

UpdateTime

The time the forecast was made.

Type: Timestamp

Errors

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

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API 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

ListTagsForResource

Returns all the tags on the specified Application Auto Scaling scalable target.

For general information about tags, including the format and syntax, see Tagging your AWS resources in the *Amazon Web Services General Reference*.

Request Syntax

```
{
   "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.

ResourceARN

Specify the ARN of the scalable target.

For example: `arn:aws:application-autoscaling:us-east-1:123456789012:scalable-target/1234abcd56ab78cd901ef1234567890ab123`

To get the ARN for a scalable target, use DescribeScalableTargets.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1011.

Pattern: `^arn:.+:application-autoscaling:.+:[0-9]+:scalable-target\/[a-zA-Z0-9-]+$`

Required: Yes

Response Syntax

```
{
   "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

A list of tags. Each tag consists of a tag key and a tag value.

Type: String to string map

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

Value 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.

ResourceNotFoundException

The specified resource doesn't exist.

ResourceName

The name of the Application Auto Scaling resource. This value is an Amazon Resource Name (ARN).

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example

The following example lists the tag key names and values that are attached to the scalable target specified by its ARN.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.ListTagsForResource
X-Amz-Date: 20230206T120729Z
User-Agent: aws-cli/2.0.0 Python/3.7.5 Windows/10 botocore/2.0.0dev4
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "ResourceARN": "arn:aws:application-autoscaling:us-west-2:123456789012:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Sample Response

```
HTTP/1.1 200 OK
x-amzn-RequestId: [request-id]
Content-Type: application/x-amz-json-1.1
Content-Length: 47
Date: Mon, 06 February 2023 12:07:29 GMT

{
    "Tags": {
        "environment": "production"
    }
}
```

See Also

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

- AWS Command Line Interface V2

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

PutScalingPolicy

Creates or updates a scaling policy for an Application Auto Scaling scalable target.

Each scalable target is identified by a service namespace, resource ID, and scalable dimension. A scaling policy applies to the scalable target identified by those three attributes. You cannot create a scaling policy until you have registered the resource as a scalable target.

Multiple scaling policies can be in force at the same time for the same scalable target. You can have one or more target tracking scaling policies, one or more step scaling policies, or both. However, there is a chance that multiple policies could conflict, instructing the scalable target to scale out or in at the same time. Application Auto Scaling gives precedence to the policy that provides the largest capacity for both scale out and scale in. For example, if one policy increases capacity by 3, another policy increases capacity by 200 percent, and the current capacity is 10, Application Auto Scaling uses the policy with the highest calculated capacity (200% of 10 = 20) and scales out to 30.

We recommend caution, however, when using target tracking scaling policies with step scaling policies because conflicts between these policies can cause undesirable behavior. For example, if the step scaling policy initiates a scale-in activity before the target tracking policy is ready to scale in, the scale-in activity will not be blocked. After the scale-in activity completes, the target tracking policy could instruct the scalable target to scale out again.

For more information, see Target tracking scaling policies, Step scaling policies, and Predictive scaling policies in the *Application Auto Scaling User Guide*.

> **Note**
>
> If a scalable target is deregistered, the scalable target is no longer available to use scaling policies. Any scaling policies that were specified for the scalable target are deleted.

Request Syntax

```
{
   "PolicyName": "string",
   "PolicyType": "string",
   "PredictiveScalingPolicyConfiguration": {
      "MaxCapacityBreachBehavior": "string",
      "MaxCapacityBuffer": number,
      "MetricSpecifications": [
```

```
{
   "CustomizedCapacityMetricSpecification": {
      "MetricDataQueries": [
         {
            "Expression": "string",
            "Id": "string",
            "Label": "string",
            "MetricStat": {
               "Metric": {
                  "Dimensions": [
                     {
                        "Name": "string",
                        "Value": "string"
                     }
                  ],
                  "MetricName": "string",
                  "Namespace": "string"
               },
               "Stat": "string",
               "Unit": "string"
            },
            "ReturnData": boolean
         }
      ]
   },
   "CustomizedLoadMetricSpecification": {
      "MetricDataQueries": [
         {
            "Expression": "string",
            "Id": "string",
            "Label": "string",
            "MetricStat": {
               "Metric": {
                  "Dimensions": [
                     {
                        "Name": "string",
                        "Value": "string"
                     }
                  ],
                  "MetricName": "string",
                  "Namespace": "string"
               },
               "Stat": "string",
               "Unit": "string"
```

```
                    },
                    "ReturnData": boolean
                }
            ]
        },
        "CustomizedScalingMetricSpecification": {
            "MetricDataQueries": [
                {
                    "Expression": "string",
                    "Id": "string",
                    "Label": "string",
                    "MetricStat": {
                        "Metric": {
                            "Dimensions": [
                                {
                                    "Name": "string",
                                    "Value": "string"
                                }
                            ],
                            "MetricName": "string",
                            "Namespace": "string"
                        },
                        "Stat": "string",
                        "Unit": "string"
                    },
                    "ReturnData": boolean
                }
            ]
        },
        "PredefinedLoadMetricSpecification": {
            "PredefinedMetricType": "string",
            "ResourceLabel": "string"
        },
        "PredefinedMetricPairSpecification": {
            "PredefinedMetricType": "string",
            "ResourceLabel": "string"
        },
        "PredefinedScalingMetricSpecification": {
            "PredefinedMetricType": "string",
            "ResourceLabel": "string"
        },
        "TargetValue": number
    }
],
```

```
      "Mode": "string",
      "SchedulingBufferTime": number
   },
   "ResourceId": "string",
   "ScalableDimension": "string",
   "ServiceNamespace": "string",
   "StepScalingPolicyConfiguration": {
      "AdjustmentType": "string",
      "Cooldown": number,
      "MetricAggregationType": "string",
      "MinAdjustmentMagnitude": number,
      "StepAdjustments": [
         {
            "MetricIntervalLowerBound": number,
            "MetricIntervalUpperBound": number,
            "ScalingAdjustment": number
         }
      ]
   },
   "TargetTrackingScalingPolicyConfiguration": {
      "CustomizedMetricSpecification": {
         "Dimensions": [
            {
               "Name": "string",
               "Value": "string"
            }
         ],
         "MetricName": "string",
         "Metrics": [
            {
               "Expression": "string",
               "Id": "string",
               "Label": "string",
               "MetricStat": {
                  "Metric": {
                     "Dimensions": [
                        {
                           "Name": "string",
                           "Value": "string"
                        }
                     ],
                     "MetricName": "string",
                     "Namespace": "string"
                  },
```

```
                    "Stat": "string",
                    "Unit": "string"
                },
                "ReturnData": boolean
            }
         ],
         "Namespace": "string",
         "Statistic": "string",
         "Unit": "string"
      },
      "DisableScaleIn": boolean,
      "PredefinedMetricSpecification": {
         "PredefinedMetricType": "string",
         "ResourceLabel": "string"
      },
      "ScaleInCooldown": number,
      "ScaleOutCooldown": number,
      "TargetValue": 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.

PolicyName

The name of the scaling policy.

You cannot change the name of a scaling policy, but you can delete the original scaling policy and create a new scaling policy with the same settings and a different name.

Type: String

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

Pattern: `\p{Print}+`

Required: Yes

PolicyType

The scaling policy type. This parameter is required if you are creating a scaling policy.

The following policy types are supported:

`TargetTrackingScaling`—Not supported for Amazon EMR.

`StepScaling`—Not supported for DynamoDB, Amazon Comprehend, Lambda, Amazon Keyspaces, Amazon MSK, Amazon ElastiCache, or Neptune.

`PredictiveScaling`—Only supported for Amazon ECS.

For more information, see Target tracking scaling policies, Step scaling policies, and Predictive scaling policies in the *Application Auto Scaling User Guide*.

Type: String

Valid Values: `StepScaling | TargetTrackingScaling | PredictiveScaling`

Required: No

PredictiveScalingPolicyConfiguration

The configuration of the predictive scaling policy.

Type: PredictiveScalingPolicyConfiguration object

Required: No

ResourceId

The identifier of the resource associated with the scaling policy. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.

- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.

- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension

The scalable dimension. This string consists of the service namespace, resource type, and scaling property.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- `sagemaker:variant:DesiredInstanceCount` - The number of EC2 instances for a SageMaker model endpoint variant.
- `custom-resource:ResourceType:Property` - The scalable dimension for a custom resource provided by your own application or service.

- `comprehend:document-classifier-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend document classification endpoint.
- `comprehend:entity-recognizer-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- `lambda:function:ProvisionedConcurrency` - The provisioned concurrency for a Lambda function.
- `cassandra:table:ReadCapacityUnits` - The provisioned read capacity for an Amazon Keyspaces table.
- `cassandra:table:WriteCapacityUnits` - The provisioned write capacity for an Amazon Keyspaces table.
- `kafka:broker-storage:VolumeSize` - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- `elasticache:cache-cluster:Nodes` - The number of nodes for an Amazon ElastiCache cache cluster.
- `elasticache:replication-group:NodeGroups` - The number of node groups for an Amazon ElastiCache replication group.
- `elasticache:replication-group:Replicas` - The number of replicas per node group for an Amazon ElastiCache replication group.
- `neptune:cluster:ReadReplicaCount` - The count of read replicas in an Amazon Neptune DB cluster.
- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-`

`resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: Yes

ServiceNamespace

The namespace of the AWS service that provides the resource. For a resource provided by your own application or service, use `custom-resource` instead.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

StepScalingPolicyConfiguration

A step scaling policy.

This parameter is required if you are creating a policy and the policy type is `StepScaling`.

Type: StepScalingPolicyConfiguration object

Required: No

TargetTrackingScalingPolicyConfiguration

A target tracking scaling policy. Includes support for predefined or customized metrics.

This parameter is required if you are creating a policy and the policy type is `TargetTrackingScaling`.

Type: TargetTrackingScalingPolicyConfiguration object

Required: No

Response Syntax

```
{
   "Alarms": [
      {
         "AlarmARN": "string",
         "AlarmName": "string"
      }
   ],
   "PolicyARN": "string"
}
```

Response Elements

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

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

Alarms

The CloudWatch alarms created for the target tracking scaling policy.

Type: Array of Alarm objects

PolicyARN

The Amazon Resource Name (ARN) of the resulting scaling policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Errors

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

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to an Application Auto Scaling resource that already has a pending update.

HTTP Status Code: 400

FailedResourceAccessException

Failed access to resources caused an exception. This exception is thrown when Application Auto Scaling is unable to retrieve the alarms associated with a scaling policy due to a client error, for example, if the role ARN specified for a scalable target does not have permission to call the CloudWatch DescribeAlarms on your behalf.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

LimitExceededException

A per-account resource limit is exceeded. For more information, see Application Auto Scaling service quotas.

HTTP Status Code: 400

ObjectNotFoundException

The specified object could not be found. For any operation that depends on the existence of a scalable target, this exception is thrown if the scalable target with the specified service namespace, resource ID, and scalable dimension does not exist. For any operation that deletes or deregisters a resource, this exception is thrown if the resource cannot be found.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example of a target tracking scaling policy

The following example applies a target tracking scaling policy to an Amazon ECS service called `web-app` in the `default` cluster. The policy keeps the average CPU utilization of the service at 75 percent, with scale-out and scale-in cooldown periods of 60 seconds. The output contains the ARNs and names of the two CloudWatch alarms created on your behalf.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.PutScalingPolicy
X-Amz-Date: 20190506T191044Z
User-Agent: aws-cli/1.10.23 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "PolicyName": "cpu75-target-tracking-scaling-policy",
    "PolicyType": "TargetTrackingScaling",
    "TargetTrackingScalingPolicyConfiguration": {
        "TargetValue": 75.0,
        "PredefinedMetricSpecification": {
            "PredefinedMetricType": "ECSServiceAverageCPUUtilization"
        },
        "ScaleOutCooldown": 60,
        "ScaleInCooldown": 60
    },
    "ServiceNamespace": "ecs",
    "ScalableDimension": "ecs:service:DesiredCount",
    "ResourceId": "service/my-cluster/my-service"
}
```

Sample Response

```
HTTP/1.1 200 OK
x-amzn-RequestId: [request-id]
Content-Type: application/x-amz-json-1.1
Content-Length: 314
Date: Fri, 06 May 2019 19:10:44 GMT

{
    "PolicyARN": "arn:aws:autoscaling:us-west-2:012345678910:scalingPolicy:6d8972f3-
efc8-437c-92d1-6270f29a66e7:resource/ecs/service/my-cluster/my-service:policyName/
cpu75-target-tracking-scaling-policy",
    "Alarms": [
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:012345678910:alarm:TargetTracking-service/my-cluster/my-service-AlarmHigh-
d4f0770c-b46e-434a-a60f-3b36d653feca",
            "AlarmName": "TargetTracking-service/my-cluster/my-service-AlarmHigh-
d4f0770c-b46e-434a-a60f-3b36d653feca"
        },
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:012345678910:alarm:TargetTracking-service/my-cluster/my-service-
AlarmLow-1b437334-d19b-4a63-a812-6c67aaf2910d",
            "AlarmName": "TargetTracking-service/my-cluster/my-service-
AlarmLow-1b437334-d19b-4a63-a812-6c67aaf2910d"
        }
    ]
}
```

Example of a step scaling policy for scale out

The following example applies a step scaling policy to an Amazon ECS service called `web-app` in the `default` cluster. The policy increases the desired count of the service by 200%, with a cooldown period of 60 seconds. The output includes the ARN for the policy, which you use to create the CloudWatch alarm.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
```

```
X-Amz-Target: AnyScaleFrontendService.PutScalingPolicy
X-Amz-Date: 20190506T191138Z
User-Agent: aws-cli/1.10.23 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "PolicyName": "my-scale-out-policy",
    "PolicyType": "StepScaling",
    "StepScalingPolicyConfiguration": {
        "AdjustmentType": "PercentChangeInCapacity",
        "Cooldown": 60,
        "MetricAggregationType": "Average",
        "StepAdjustments": [
            {
                "ScalingAdjustment": 200,
                "MetricIntervalLowerBound": 0
            }
        ]
    },
    "ServiceNamespace": "ecs",
    "ScalableDimension": "ecs:service:DesiredCount",
    "ResourceId": "service/my-cluster/my-service"
}
```

Sample Response

```
HTTP/1.1 200 OK
x-amzn-RequestId: [request-id]
Content-Type: application/x-amz-json-1.1
Content-Length: 175
Date: Fri, 06 May 2019 19:11:38 GMT

{
    "PolicyARN": "arn:aws:autoscaling:us-west-2:012345678910:scalingPolicy:ac542982-
cbeb-4294-891c-a5a941dfa787:resource/ecs/service/my-cluster/my-service:policyName/my-
scale-out-policy"
}
```

Example of a step scaling policy for scale in

The following example applies a step scaling policy to the same Amazon ECS service as in the preceding example. The policy has two step adjustments that decrease the desired count of the

service by 25% or 50%, depending on the size of the alarm breach, with a cooldown period of 120 seconds. The output includes the ARN for the policy, which you use to create the CloudWatch alarm.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.PutScalingPolicy
X-Amz-Date: 20190506T191152Z
User-Agent: aws-cli/1.10.23 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "PolicyName": "my-scale-in-policy",
    "PolicyType": "StepScaling",
    "StepScalingPolicyConfiguration": {
        "AdjustmentType": "PercentChangeInCapacity",
        "Cooldown": 120,
        "MetricAggregationType": "Average",
        "MinAdjustmentMagnitude": 1,
        "StepAdjustments": [
            {
                "ScalingAdjustment": -25,
                "MetricIntervalLowerBound": -15,
                "MetricIntervalUpperBound": 0
             },
            {
                "ScalingAdjustment": -50,
                "MetricIntervalUpperBound": -15
            }
        ]
    },
    "ServiceNamespace": "ecs",
    "ScalableDimension": "ecs:service:DesiredCount",
    "ResourceId": "service/my-cluster/my-service"
}
```

Sample Response

```
HTTP/1.1 200 OK
x-amzn-RequestId: [request-id]
Content-Type: application/x-amz-json-1.1
Content-Length: 174
Date: Fri, 06 May 2019 19:11:52 GMT

{
    "PolicyARN": "arn:aws:autoscaling:us-west-2:012345678910:scalingPolicy:6d8972f3-efc8-437c-92d1-6270f29a66e7:resource/ecs/service/my-cluster/my-service:policyName/my-scale-in-policy"
}
```

See Also

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

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

PutScheduledAction

Creates or updates a scheduled action for an Application Auto Scaling scalable target.

Each scalable target is identified by a service namespace, resource ID, and scalable dimension. A scheduled action applies to the scalable target identified by those three attributes. You cannot create a scheduled action until you have registered the resource as a scalable target.

When you specify start and end times with a recurring schedule using a cron expression or rates, they form the boundaries for when the recurring action starts and stops.

To update a scheduled action, specify the parameters that you want to change. If you don't specify start and end times, the old values are deleted.

For more information, see Scheduled scaling in the *Application Auto Scaling User Guide*.

> **Note**
>
> If a scalable target is deregistered, the scalable target is no longer available to run scheduled actions. Any scheduled actions that were specified for the scalable target are deleted.

Request Syntax

```
{
   "EndTime": number,
   "ResourceId": "string",
   "ScalableDimension": "string",
   "ScalableTargetAction": {
      "MaxCapacity": number,
      "MinCapacity": number
   },
   "Schedule": "string",
   "ScheduledActionName": "string",
   "ServiceNamespace": "string",
   "StartTime": number,
   "Timezone": "string"
}
```

Request Parameters

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

The request accepts the following data in JSON format.

EndTime

The date and time for the recurring schedule to end, in UTC.

Type: Timestamp

Required: No

ResourceId

The identifier of the resource associated with the scheduled action. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The

unique identifier is defined by the service provider. More information is available in our GitHub repository.

- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension

The scalable dimension. This string consists of the service namespace, resource type, and scaling property.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- `sagemaker:variant:DesiredInstanceCount` - The number of EC2 instances for a SageMaker model endpoint variant.
- `custom-resource:ResourceType:Property` - The scalable dimension for a custom resource provided by your own application or service.
- `comprehend:document-classifier-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend document classification endpoint.
- `comprehend:entity-recognizer-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- `lambda:function:ProvisionedConcurrency` - The provisioned concurrency for a Lambda function.
- `cassandra:table:ReadCapacityUnits` - The provisioned read capacity for an Amazon Keyspaces table.

- `cassandra:table:WriteCapacityUnits` - The provisioned write capacity for an Amazon Keyspaces table.
- `kafka:broker-storage:VolumeSize` - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- `elasticache:cache-cluster:Nodes` - The number of nodes for an Amazon ElastiCache cache cluster.
- `elasticache:replication-group:NodeGroups` - The number of node groups for an Amazon ElastiCache replication group.
- `elasticache:replication-group:Replicas` - The number of replicas per node group for an Amazon ElastiCache replication group.
- `neptune:cluster:ReadReplicaCount` - The count of read replicas in an Amazon Neptune DB cluster.
- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency`

| sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions

Required: Yes

ScalableTargetAction

The new minimum and maximum capacity. You can set both values or just one. At the scheduled time, if the current capacity is below the minimum capacity, Application Auto Scaling scales out to the minimum capacity. If the current capacity is above the maximum capacity, Application Auto Scaling scales in to the maximum capacity.

Type: ScalableTargetAction object

Required: No

Schedule

The schedule for this action. The following formats are supported:

- At expressions - "at(*yyyy-mm-dd*T*hh*:*mm*:*ss*)"
- Rate expressions - "rate(*value unit*)"
- Cron expressions - "cron(*fields*)"

At expressions are useful for one-time schedules. Cron expressions are useful for scheduled actions that run periodically at a specified date and time, and rate expressions are useful for scheduled actions that run at a regular interval.

At and cron expressions use Universal Coordinated Time (UTC) by default.

The cron format consists of six fields separated by white spaces: [Minutes] [Hours] [Day_of_Month] [Month] [Day_of_Week] [Year].

For rate expressions, *value* is a positive integer and *unit* is minute | minutes | hour | hours | day | days.

For more information, see Schedule recurring scaling actions using cron expressions in the *Application Auto Scaling User Guide*.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: [\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*

Required: No

ScheduledActionName

The name of the scheduled action. This name must be unique among all other scheduled actions on the specified scalable target.

Type: String

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

Pattern: `(?!((^[ ]+.*)|(.*([\u0000-\u001f]|[\u007f-\u009f]|[:/|])+.*)|(.*[ ]+$))).+`

Required: Yes

ServiceNamespace

The namespace of the AWS service that provides the resource. For a resource provided by your own application or service, use `custom-resource` instead.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

StartTime

The date and time for this scheduled action to start, in UTC.

Type: Timestamp

Required: No

Timezone

Specifies the time zone used when setting a scheduled action by using an at or cron expression. If a time zone is not provided, UTC is used by default.

Valid values are the canonical names of the IANA time zones supported by Joda-Time (such as `Etc/GMT+9` or `Pacific/Tahiti`). For more information, see https://www.joda.org/joda-time/timezones.html.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

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.

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to an Application Auto Scaling resource that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

LimitExceededException

A per-account resource limit is exceeded. For more information, see Application Auto Scaling service quotas.

HTTP Status Code: 400

ObjectNotFoundException

The specified object could not be found. For any operation that depends on the existence of a scalable target, this exception is thrown if the scalable target with the specified service namespace, resource ID, and scalable dimension does not exist. For any operation that deletes or deregisters a resource, this exception is thrown if the resource cannot be found.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example of a scheduled action

This example adds a scheduled action to a DynamoDB table called `TestTable` to scale out on a recurring schedule. On the specified schedule (every day at 12:15pm UTC), if the current capacity is below the value specified for `MinCapacity`, Application Auto Scaling scales out to the value specified by `MinCapacity`.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.PutScheduledAction
X-Amz-Date: 20190506T191138Z
User-Agent: aws-cli/1.10.23 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "ScheduledActionName": "my-recurring-action",
    "Schedule": "cron(15 12 * * ? *)",
    "ScalableTargetAction": {
        "MinCapacity": 6
    },
    "ServiceNamespace": "dynamodb",
    "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
```

```
    "ResourceId": "table/TestTable"
}
```

See Also

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

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

RegisterScalableTarget

Registers or updates a scalable target, which is the resource that you want to scale.

Scalable targets are uniquely identified by the combination of resource ID, scalable dimension, and namespace, which represents some capacity dimension of the underlying service.

When you register a new scalable target, you must specify values for the minimum and maximum capacity. If the specified resource is not active in the target service, this operation does not change the resource's current capacity. Otherwise, it changes the resource's current capacity to a value that is inside of this range.

If you add a scaling policy, current capacity is adjustable within the specified range when scaling starts. Application Auto Scaling scaling policies will not scale capacity to values that are outside of the minimum and maximum range.

After you register a scalable target, you do not need to register it again to use other Application Auto Scaling operations. To see which resources have been registered, use DescribeScalableTargets. You can also view the scaling policies for a service namespace by using DescribeScalableTargets. If you no longer need a scalable target, you can deregister it by using DeregisterScalableTarget.

To update a scalable target, specify the parameters that you want to change. Include the parameters that identify the scalable target: resource ID, scalable dimension, and namespace. Any parameters that you don't specify are not changed by this update request.

Note

If you call the `RegisterScalableTarget` API operation to create a scalable target, there might be a brief delay until the operation achieves eventual consistency. You might become aware of this brief delay if you get unexpected errors when performing sequential operations. The typical strategy is to retry the request, and some AWS SDKs include automatic backoff and retry logic.
If you call the `RegisterScalableTarget` API operation to update an existing scalable target, Application Auto Scaling retrieves the current capacity of the resource. If it's below the minimum capacity or above the maximum capacity, Application Auto Scaling adjusts the capacity of the scalable target to place it within these bounds, even if you don't include the `MinCapacity` or `MaxCapacity` request parameters.

Request Syntax

```
{
   "MaxCapacity": number,
   "MinCapacity": number,
   "ResourceId": "string",
   "RoleARN": "string",
   "ScalableDimension": "string",
   "ServiceNamespace": "string",
   "SuspendedState": {
      "DynamicScalingInSuspended": boolean,
      "DynamicScalingOutSuspended": boolean,
      "ScheduledScalingSuspended": boolean
   },
   "Tags": {
      "string" : "string"
   }
}
```

Request Parameters

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

The request accepts the following data in JSON format.

MaxCapacity

The maximum value that you plan to scale out to. When a scaling policy is in effect, Application Auto Scaling can scale out (expand) as needed to the maximum capacity limit in response to changing demand. This property is required when registering a new scalable target.

Although you can specify a large maximum capacity, note that service quotas might impose lower limits. Each service has its own default quotas for the maximum capacity of the resource. If you want to specify a higher limit, you can request an increase. For more information, consult the documentation for that service. For information about the default quotas for each service, see Service endpoints and quotas in the *Amazon Web Services General Reference*.

Type: Integer

Required: No

MinCapacity

The minimum value that you plan to scale in to. When a scaling policy is in effect, Application Auto Scaling can scale in (contract) as needed to the minimum capacity limit in response to changing demand. This property is required when registering a new scalable target.

For the following resources, the minimum value allowed is 0.

- AppStream 2.0 fleets
- Aurora DB clusters
- ECS services
- EMR clusters
- Lambda provisioned concurrency
- SageMaker endpoint variants
- SageMaker inference components
- SageMaker serverless endpoint provisioned concurrency
- Spot Fleets
- custom resources

It's strongly recommended that you specify a value greater than 0. A value greater than 0 means that data points are continuously reported to CloudWatch that scaling policies can use to scale on a metric like average CPU utilization.

For all other resources, the minimum allowed value depends on the type of resource that you are using. If you provide a value that is lower than what a resource can accept, an error occurs. In which case, the error message will provide the minimum value that the resource can accept.

Type: Integer

Required: No

ResourceId

The identifier of the resource that is associated with the scalable target. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.

- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.

- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

RoleARN

This parameter is required for services that do not support service-linked roles (such as Amazon EMR), and it must specify the ARN of an IAM role that allows Application Auto Scaling to modify the scalable target on your behalf.

If the service supports service-linked roles, Application Auto Scaling uses a service-linked role, which it creates if it does not yet exist. For more information, see How Application Auto Scaling works with IAM.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

ScalableDimension

The scalable dimension associated with the scalable target. This string consists of the service namespace, resource type, and scaling property.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- `sagemaker:variant:DesiredInstanceCount` - The number of EC2 instances for a SageMaker model endpoint variant.
- `custom-resource:ResourceType:Property` - The scalable dimension for a custom resource provided by your own application or service.
- `comprehend:document-classifier-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend document classification endpoint.
- `comprehend:entity-recognizer-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- `lambda:function:ProvisionedConcurrency` - The provisioned concurrency for a Lambda function.
- `cassandra:table:ReadCapacityUnits` - The provisioned read capacity for an Amazon Keyspaces table.
- `cassandra:table:WriteCapacityUnits` - The provisioned write capacity for an Amazon Keyspaces table.

- `kafka:broker-storage:VolumeSize` - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- `elasticache:cache-cluster:Nodes` - The number of nodes for an Amazon ElastiCache cache cluster.
- `elasticache:replication-group:NodeGroups` - The number of node groups for an Amazon ElastiCache replication group.
- `elasticache:replication-group:Replicas` - The number of replicas per node group for an Amazon ElastiCache replication group.
- `neptune:cluster:ReadReplicaCount` - The count of read replicas in an Amazon Neptune DB cluster.
- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: Yes

ServiceNamespace

The namespace of the AWS service that provides the resource. For a resource provided by your own application or service, use `custom-resource` instead.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

SuspendedState

An embedded object that contains attributes and attribute values that are used to suspend and resume automatic scaling. Setting the value of an attribute to `true` suspends the specified scaling activities. Setting it to `false` (default) resumes the specified scaling activities.

Suspension Outcomes

- For `DynamicScalingInSuspended`, while a suspension is in effect, all scale-in activities that are triggered by a scaling policy are suspended.
- For `DynamicScalingOutSuspended`, while a suspension is in effect, all scale-out activities that are triggered by a scaling policy are suspended.
- For `ScheduledScalingSuspended`, while a suspension is in effect, all scaling activities that involve scheduled actions are suspended.

For more information, see Suspend and resume scaling in the *Application Auto Scaling User Guide*.

Type: SuspendedState object

Required: No

Tags

Assigns one or more tags to the scalable target. Use this parameter to tag the scalable target when it is created. To tag an existing scalable target, use the TagResource operation.

Each tag consists of a tag key and a tag value. Both the tag key and the tag value are required. You cannot have more than one tag on a scalable target with the same tag key.

Use tags to control access to a scalable target. For more information, see Tagging support for Application Auto Scaling in the *Application Auto Scaling User Guide*.

Type: String to string map

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

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

Required: No

Response Syntax

```
{
   "ScalableTargetARN": "string"
}
```

Response Elements

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

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

ScalableTargetARN

The ARN of the scalable target.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Errors

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

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to an Application Auto Scaling resource that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

LimitExceededException

A per-account resource limit is exceeded. For more information, see Application Auto Scaling service quotas.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example

The following example registers an Amazon ECS service with Application Auto Scaling.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.RegisterScalableTarget
X-Amz-Date: 20190506T182145Z
User-Agent: aws-cli/1.10.23 Python/2.7.11 Darwin/15.4.0 botocore/1.4.8
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS
```

```
{
    "ScalableDimension": "ecs:service:DesiredCount",
    "ResourceId": "service/my-cluster/my-service",
    "ServiceNamespace": "ecs",
    "MinCapacity": 1,
    "MaxCapacity": 10,
    "Tags": {
        "environment" : "production"
   }
}
```

Sample Response

```
HTTP/1.1 200 OK
x-amzn-RequestId: [request-id]
Content-Type: application/x-amz-json-1.1
Content-Length: 142
Date: Fri, 20 Mar 2023 12:10:44 GMT

{
    "ScalableTargetARN": "arn:aws:application-autoscaling:us-
east-1:123456789012:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

See Also

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

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

Adds or edits tags on an Application Auto Scaling scalable target.

Each tag consists of a tag key and a tag value, which are both case-sensitive strings. To add a tag, specify a new tag key and a tag value. To edit a tag, specify an existing tag key and a new tag value.

You can use this operation to tag an Application Auto Scaling scalable target, but you cannot tag a scaling policy or scheduled action.

You can also add tags to an Application Auto Scaling scalable target while creating it (`RegisterScalableTarget`).

For general information about tags, including the format and syntax, see Tagging your AWS resources in the *Amazon Web Services General Reference*.

Use tags to control access to a scalable target. For more information, see Tagging support for Application Auto Scaling in the *Application Auto Scaling User Guide*.

Request Syntax

```
{
   "ResourceARN": "string",
   "Tags": {
      "string" : "string"
   }
}
```

Request 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

Identifies the Application Auto Scaling scalable target that you want to apply tags to.

For example: `arn:aws:application-autoscaling:us-east-1:123456789012:scalable-target/1234abcd56ab78cd901ef1234567890ab123`

To get the ARN for a scalable target, use DescribeScalableTargets.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1011.

Pattern: `^arn:.+:application-autoscaling:.+:[0-9]+:scalable-target\/[a-zA-Z0-9-]+$`

Required: Yes

Tags

The tags assigned to the resource. A tag is a label that you assign to an AWS resource.

Each tag consists of a tag key and a tag value.

You cannot have more than one tag on an Application Auto Scaling scalable target with the same tag key. If you specify an existing tag key with a different tag value, Application Auto Scaling replaces the current tag value with the specified one.

For information about the rules that apply to tag keys and tag values, see User-defined tag restrictions in the *AWS Billing User Guide*.

Type: String to string map

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

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

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.

ResourceNotFoundException

The specified resource doesn't exist.

ResourceName

The name of the Application Auto Scaling resource. This value is an Amazon Resource Name (ARN).

HTTP Status Code: 400

TooManyTagsException

The request contains too many tags. Try the request again with fewer tags.

ResourceName

The name of the Application Auto Scaling resource. This value is an Amazon Resource Name (ARN).

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example

The following example adds a tag with the key name "environment" and the value "production" to the scalable target specified by its ARN.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
```

```
X-Amz-Target: AnyScaleFrontendService.TagResource
X-Amz-Date: 20230506T182145Z
User-Agent: aws-cli/2.0.0 Python/3.7.5 Windows/10 botocore/2.0.0dev4
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "ResourceARN": "arn:aws:application-autoscaling:us-west-2:123456789012:scalable-
target/1234abcd56ab78cd901ef1234567890ab123",
    "Tags": {
        "environment": "production"
    }
}
```

See Also

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

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

Deletes tags from an Application Auto Scaling scalable target. To delete a tag, specify the tag key and the Application Auto Scaling scalable target.

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

Identifies the Application Auto Scaling scalable target from which to remove tags.

For example: `arn:aws:application-autoscaling:us-east-1:123456789012:scalable-target/1234abcd56ab78cd901ef1234567890ab123`

To get the ARN for a scalable target, use DescribeScalableTargets.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1011.

Pattern: `^arn:.+:application-autoscaling:.+:[0-9]+:scalable-target\/[a-zA-Z0-9-]+$`

Required: Yes

TagKeys

One or more tag keys. Specify only the tag keys, not the tag values.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 200 items.

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

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.

ResourceNotFoundException

The specified resource doesn't exist.

ResourceName

The name of the Application Auto Scaling resource. This value is an Amazon Resource Name (ARN).

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the available parameters for the API request.

HTTP Status Code: 400

Examples

If you plan to create requests manually, you must replace the Authorization header contents in the examples (AUTHPARAMS) with a signature. For more information, see Signing AWS API requests in the *IAM User Guide*. If you plan to use the AWS CLI or one of the AWS SDKs, these tools sign the requests for you.

Example

The following example removes the tag pair with the key name "environment" from the scalable target specified by its ARN.

Sample Request

```
POST / HTTP/1.1
Host: application-autoscaling.us-west-2.amazonaws.com
Accept-Encoding: identity
Content-Length: [content-length]
X-Amz-Target: AnyScaleFrontendService.UntagResource
X-Amz-Date: 20230506T211829Z
User-Agent: aws-cli/2.0.0 Python/3.7.5 Windows/10 botocore/2.0.0dev4
Content-Type: application/x-amz-json-1.1
Authorization: AUTHPARAMS

{
    "ResourceARN": "arn:aws:application-autoscaling:us-west-2:123456789012:scalable-
target/1234abcd56ab78cd901ef1234567890ab123",
    "TagKeys": [ "environment" ]
}
```

See Also

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

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java 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 Application Auto Scaling API contains several data types that various actions use. This section describes each data type in detail.

> **Note**
>
> The order of each element in a data type structure is not guaranteed. Applications should not assume a particular order.

The following data types are supported:

- Alarm
- CapacityForecast
- CustomizedMetricSpecification
- LoadForecast
- MetricDimension
- NotScaledReason
- PredefinedMetricSpecification
- PredictiveScalingCustomizedMetricSpecification
- PredictiveScalingMetric
- PredictiveScalingMetricDataQuery
- PredictiveScalingMetricDimension
- PredictiveScalingMetricSpecification
- PredictiveScalingMetricStat
- PredictiveScalingPolicyConfiguration
- PredictiveScalingPredefinedLoadMetricSpecification
- PredictiveScalingPredefinedMetricPairSpecification
- PredictiveScalingPredefinedScalingMetricSpecification
- ScalableTarget
- ScalableTargetAction
- ScalingActivity

- ScalingPolicy
- ScheduledAction
- StepAdjustment
- StepScalingPolicyConfiguration
- SuspendedState
- TargetTrackingMetric
- TargetTrackingMetricDataQuery
- TargetTrackingMetricDimension
- TargetTrackingMetricStat
- TargetTrackingScalingPolicyConfiguration

Alarm

Represents a CloudWatch alarm associated with a scaling policy.

Contents

AlarmARN

The Amazon Resource Name (ARN) of the alarm.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

AlarmName

The name of the alarm.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\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

CapacityForecast

A `GetPredictiveScalingForecast` call returns the capacity forecast for a predictive scaling policy. This structure includes the data points for that capacity forecast, along with the timestamps of those data points.

Contents

Timestamps

The timestamps for the data points, in UTC format.

Type: Array of timestamps

Required: Yes

Values

The values of the data points.

Type: Array of doubles

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

CustomizedMetricSpecification

Represents a CloudWatch metric of your choosing for a target tracking scaling policy to use with Application Auto Scaling.

For information about the available metrics for a service, see AWS services that publish CloudWatch metrics in the *Amazon CloudWatch User Guide*.

To create your customized metric specification:

- Add values for each required parameter from CloudWatch. You can use an existing metric, or a new metric that you create. To use your own metric, you must first publish the metric to CloudWatch. For more information, see Publish custom metrics in the *Amazon CloudWatch User Guide*.
- Choose a metric that changes proportionally with capacity. The value of the metric should increase or decrease in inverse proportion to the number of capacity units. That is, the value of the metric should decrease when capacity increases, and increase when capacity decreases.

For more information about the CloudWatch terminology below, see Amazon CloudWatch concepts in the *Amazon CloudWatch User Guide*.

Contents

Dimensions

The dimensions of the metric.

Conditional: If you published your metric with dimensions, you must specify the same dimensions in your scaling policy.

Type: Array of MetricDimension objects

Required: No

MetricName

The name of the metric. To get the exact metric name, namespace, and dimensions, inspect the Metric object that's returned by a call to ListMetrics.

Type: String

Required: No

Metrics

The metrics to include in the target tracking scaling policy, as a metric data query. This can include both raw metric and metric math expressions.

Type: Array of TargetTrackingMetricDataQuery objects

Required: No

Namespace

The namespace of the metric.

Type: String

Required: No

Statistic

The statistic of the metric.

Type: String

Valid Values: `Average | Minimum | Maximum | SampleCount | Sum`

Required: No

Unit

The unit of the metric. For a complete list of the units that CloudWatch supports, see the MetricDatum data type in the *Amazon CloudWatch API Reference*.

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

LoadForecast

A `GetPredictiveScalingForecast` call returns the load forecast for a predictive scaling policy. This structure includes the data points for that load forecast, along with the timestamps of those data points and the metric specification.

Contents

MetricSpecification

The metric specification for the load forecast.

Type: PredictiveScalingMetricSpecification object

Required: Yes

Timestamps

The timestamps for the data points, in UTC format.

Type: Array of timestamps

Required: Yes

Values

The values of the data points.

Type: Array of doubles

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

MetricDimension

Describes the dimension names and values associated with a metric.

Contents

Name

The name of the dimension.

Type: String

Required: Yes

Value

The value of the dimension.

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

NotScaledReason

Describes the reason for an activity that isn't scaled (*not scaled activity*), in machine-readable format. For help interpreting the not scaled reason details, see Scaling activities for Application Auto Scaling in the *Application Auto Scaling User Guide*.

Contents

Code

A code that represents the reason for not scaling.

Valid values:

- AutoScalingAnticipatedFlapping
- TargetServicePutResourceAsUnscalable
- AlreadyAtMaxCapacity
- AlreadyAtMinCapacity
- AlreadyAtDesiredCapacity

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

CurrentCapacity

The current capacity.

Type: Integer

Required: No

MaxCapacity

The maximum capacity.

Type: Integer

Required: No

MinCapacity

The minimum capacity.

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

PredefinedMetricSpecification

Represents a predefined metric for a target tracking scaling policy to use with Application Auto Scaling.

For more information, Predefined metrics for target tracking scaling policies in the *Application Auto Scaling User Guide*.

Contents

PredefinedMetricType

The metric type. The `ALBRequestCountPerTarget` metric type applies only to Spot Fleets and ECS services.

Type: String

Valid Values: `DynamoDBReadCapacityUtilization | DynamoDBWriteCapacityUtilization | ALBRequestCountPerTarget | RDSReaderAverageCPUUtilization | RDSReaderAverageDatabaseConnections | EC2SpotFleetRequestAverageCPUUtilization | EC2SpotFleetRequestAverageNetworkIn | EC2SpotFleetRequestAverageNetworkOut | SageMakerVariantInvocationsPerInstance | ECSServiceAverageCPUUtilization | ECSServiceAverageMemoryUtilization | AppStreamAverageCapacityUtilization | ComprehendInferenceUtilization | LambdaProvisionedConcurrencyUtilization | CassandraReadCapacityUtilization | CassandraWriteCapacityUtilization | KafkaBrokerStorageUtilization | ElastiCacheEngineCPUUtilization | ElastiCacheDatabaseMemoryUsagePercentage | ElastiCachePrimaryEngineCPUUtilization | ElastiCacheReplicaEngineCPUUtilization | ElastiCacheDatabaseMemoryUsageCountedForEvictPercentage | NeptuneReaderAverageCPUUtilization | SageMakerVariantProvisionedConcurrencyUtilization | ElastiCacheDatabaseCapacityUsageCountedForEvictPercentage | SageMakerInferenceComponentInvocationsPerCopy | WorkSpacesAverageUserSessionsCapacityUtilization |`

`SageMakerInferenceComponentConcurrentRequestsPerCopyHighResolution | SageMakerVariantConcurrentRequestsPerModelHighResolution`

Required: Yes

ResourceLabel

Identifies the resource associated with the metric type. You can't specify a resource label unless the metric type is `ALBRequestCountPerTarget` and there is a target group attached to the Spot Fleet or ECS service.

You create the resource label by appending the final portion of the load balancer ARN and the final portion of the target group ARN into a single value, separated by a forward slash (/). The format of the resource label is:

`app/my-alb/778d41231b141a0f/targetgroup/my-alb-target-group/943f017f100becff`.

Where:

- app/<load-balancer-name>/<load-balancer-id> is the final portion of the load balancer ARN
- targetgroup/<target-group-name>/<target-group-id> is the final portion of the target group ARN.

To find the ARN for an Application Load Balancer, use the DescribeLoadBalancers API operation. To find the ARN for the target group, use the DescribeTargetGroups API operation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1023.

Required: No

See Also

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

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PredictiveScalingCustomizedMetricSpecification

Represents a CloudWatch metric of your choosing for a predictive scaling policy.

Contents

MetricDataQueries

One or more metric data queries to provide data points for a metric specification.

Type: Array of PredictiveScalingMetricDataQuery 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

PredictiveScalingMetric

Describes the scaling metric.

Contents

Dimensions

Describes the dimensions of the metric.

Type: Array of PredictiveScalingMetricDimension objects

Required: No

MetricName

The name of the metric.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

Namespace

The namespace of the metric.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

See Also

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

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PredictiveScalingMetricDataQuery

The metric data to return. Also defines whether this call is returning data for one metric only, or whether it is performing a math expression on the values of returned metric statistics to create a new time series. A time series is a series of data points, each of which is associated with a timestamp.

Contents

Id

A short name that identifies the object's results in the response. This name must be unique among all `MetricDataQuery` objects specified for a single scaling policy. If you are performing math expressions on this set of data, this name represents that data and can serve as a variable in the mathematical expression. The valid characters are letters, numbers, and underscores. The first character must be a lowercase letter.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

Expression

The math expression to perform on the returned data, if this object is performing a math expression. This expression can use the `Id` of the other metrics to refer to those metrics, and can also use the `Id` of other expressions to use the result of those expressions.

Conditional: Within each `MetricDataQuery` object, you must specify either `Expression` or `MetricStat`, but not both.

Type: String

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

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

Label

A human-readable label for this metric or expression. This is especially useful if this is a math expression, so that you know what the value represents.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

MetricStat

Information about the metric data to return.

Conditional: Within each `MetricDataQuery` object, you must specify either `Expression` or `MetricStat`, but not both.

Type: PredictiveScalingMetricStat object

Required: No

ReturnData

Indicates whether to return the timestamps and raw data values of this metric.

If you use any math expressions, specify `true` for this value for only the final math expression that the metric specification is based on. You must specify `false` for `ReturnData` for all the other metrics and expressions used in the metric specification.

If you are only retrieving metrics and not performing any math expressions, do not specify anything for `ReturnData`. This sets it to its default (`true`).

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

PredictiveScalingMetricDimension

Describes the dimension of a metric.

Contents

Name

The name of the dimension.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

Value

The value of the dimension.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\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

PredictiveScalingMetricSpecification

This structure specifies the metrics and target utilization settings for a predictive scaling policy.

You must specify either a metric pair, or a load metric and a scaling metric individually. Specifying a metric pair instead of individual metrics provides a simpler way to configure metrics for a scaling policy. You choose the metric pair, and the policy automatically knows the correct sum and average statistics to use for the load metric and the scaling metric.

Contents

TargetValue

Specifies the target utilization.

Type: Double

Required: Yes

CustomizedCapacityMetricSpecification

The customized capacity metric specification.

Type: PredictiveScalingCustomizedMetricSpecification object

Required: No

CustomizedLoadMetricSpecification

The customized load metric specification.

Type: PredictiveScalingCustomizedMetricSpecification object

Required: No

CustomizedScalingMetricSpecification

The customized scaling metric specification.

Type: PredictiveScalingCustomizedMetricSpecification object

Required: No

PredefinedLoadMetricSpecification

The predefined load metric specification.

Type: PredictiveScalingPredefinedLoadMetricSpecification object

Required: No

PredefinedMetricPairSpecification

The predefined metric pair specification that determines the appropriate scaling metric and load metric to use.

Type: PredictiveScalingPredefinedMetricPairSpecification object

Required: No

PredefinedScalingMetricSpecification

The predefined scaling metric specification.

Type: PredictiveScalingPredefinedScalingMetricSpecification 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

PredictiveScalingMetricStat

This structure defines the CloudWatch metric to return, along with the statistic and unit.

Contents

Metric

The CloudWatch metric to return, including the metric name, namespace, and dimensions. To get the exact metric name, namespace, and dimensions, inspect the Metric object that is returned by a call to ListMetrics.

Type: PredictiveScalingMetric object

Required: Yes

Stat

The statistic to return. It can include any CloudWatch statistic or extended statistic. For a list of valid values, see the table in Statistics in the *Amazon CloudWatch User Guide*.

The most commonly used metrics for predictive scaling are `Average` and `Sum`.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

Unit

The unit to use for the returned data points. For a complete list of the units that CloudWatch supports, see the MetricDatum data type in the *Amazon CloudWatch API Reference*.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1023.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

See Also

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

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PredictiveScalingPolicyConfiguration

Represents a predictive scaling policy configuration. Predictive scaling is supported on Amazon ECS services.

Contents

MetricSpecifications

This structure includes the metrics and target utilization to use for predictive scaling.

This is an array, but we currently only support a single metric specification. That is, you can specify a target value and a single metric pair, or a target value and one scaling metric and one load metric.

Type: Array of PredictiveScalingMetricSpecification objects

Required: Yes

MaxCapacityBreachBehavior

Defines the behavior that should be applied if the forecast capacity approaches or exceeds the maximum capacity. Defaults to `HonorMaxCapacity` if not specified.

Type: String

Valid Values: `HonorMaxCapacity | IncreaseMaxCapacity`

Required: No

MaxCapacityBuffer

The size of the capacity buffer to use when the forecast capacity is close to or exceeds the maximum capacity. The value is specified as a percentage relative to the forecast capacity. For example, if the buffer is 10, this means a 10 percent buffer, such that if the forecast capacity is 50, and the maximum capacity is 40, then the effective maximum capacity is 55.

Required if the `MaxCapacityBreachBehavior` property is set to `IncreaseMaxCapacity`, and cannot be used otherwise.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 100.

Required: No

Mode

The predictive scaling mode. Defaults to `ForecastOnly` if not specified.

Type: String

Valid Values: `ForecastOnly | ForecastAndScale`

Required: No

SchedulingBufferTime

The amount of time, in seconds, that the start time can be advanced.

The value must be less than the forecast interval duration of 3600 seconds (60 minutes). Defaults to 300 seconds if not specified.

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

PredictiveScalingPredefinedLoadMetricSpecification

Describes a load metric for a predictive scaling policy.

When returned in the output of `DescribePolicies`, it indicates that a predictive scaling policy uses individually specified load and scaling metrics instead of a metric pair.

The following predefined metrics are available for predictive scaling:

- `ECSServiceAverageCPUUtilization`
- `ECSServiceAverageMemoryUtilization`
- `ECSServiceCPUUtilization`
- `ECSServiceMemoryUtilization`
- `ECSServiceTotalCPUUtilization`
- `ECSServiceTotalMemoryUtilization`
- `ALBRequestCount`
- `ALBRequestCountPerTarget`
- `TotalALBRequestCount`

Contents

PredefinedMetricType

The metric type.

Type: String

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

Required: Yes

ResourceLabel

A label that uniquely identifies a target group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1023.

Required: No

See Also

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

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PredictiveScalingPredefinedMetricPairSpecification

Represents a metric pair for a predictive scaling policy.

The following predefined metrics are available for predictive scaling:

- `ECSServiceAverageCPUUtilization`
- `ECSServiceAverageMemoryUtilization`
- `ECSServiceCPUUtilization`
- `ECSServiceMemoryUtilization`
- `ECSServiceTotalCPUUtilization`
- `ECSServiceTotalMemoryUtilization`
- `ALBRequestCount`
- `ALBRequestCountPerTarget`
- `TotalALBRequestCount`

Contents

PredefinedMetricType

Indicates which metrics to use. There are two different types of metrics for each metric type: one is a load metric and one is a scaling metric.

Type: String

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

Required: Yes

ResourceLabel

A label that uniquely identifies a specific target group from which to determine the total and average request count.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1023.

Required: No

See Also

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

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PredictiveScalingPredefinedScalingMetricSpecification

Describes a scaling metric for a predictive scaling policy.

When returned in the output of `DescribePolicies`, it indicates that a predictive scaling policy uses individually specified load and scaling metrics instead of a metric pair.

The following predefined metrics are available for predictive scaling:

- `ECSServiceAverageCPUUtilization`
- `ECSServiceAverageMemoryUtilization`
- `ECSServiceCPUUtilization`
- `ECSServiceMemoryUtilization`
- `ECSServiceTotalCPUUtilization`
- `ECSServiceTotalMemoryUtilization`
- `ALBRequestCount`
- `ALBRequestCountPerTarget`
- `TotalALBRequestCount`

Contents

PredefinedMetricType

The metric type.

Type: String

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

Required: Yes

ResourceLabel

A label that uniquely identifies a specific target group from which to determine the average request count.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1023.

Required: No

See Also

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

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ScalableTarget

Represents a scalable target.

Contents

CreationTime

The Unix timestamp for when the scalable target was created.

Type: Timestamp

Required: Yes

MaxCapacity

The maximum value to scale to in response to a scale-out activity.

Type: Integer

Required: Yes

MinCapacity

The minimum value to scale to in response to a scale-in activity.

Type: Integer

Required: Yes

ResourceId

The identifier of the resource associated with the scalable target. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.

- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.

- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

RoleARN

The ARN of an IAM role that allows Application Auto Scaling to modify the scalable target on your behalf.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension

The scalable dimension associated with the scalable target. This string consists of the service namespace, resource type, and scaling property.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.

- dynamodb:table:WriteCapacityUnits - The provisioned write capacity for a DynamoDB table.
- dynamodb:index:ReadCapacityUnits - The provisioned read capacity for a DynamoDB global secondary index.
- dynamodb:index:WriteCapacityUnits - The provisioned write capacity for a DynamoDB global secondary index.
- rds:cluster:ReadReplicaCount - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- sagemaker:variant:DesiredInstanceCount - The number of EC2 instances for a SageMaker model endpoint variant.
- custom-resource:ResourceType:Property - The scalable dimension for a custom resource provided by your own application or service.
- comprehend:document-classifier-endpoint:DesiredInferenceUnits - The number of inference units for an Amazon Comprehend document classification endpoint.
- comprehend:entity-recognizer-endpoint:DesiredInferenceUnits - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- lambda:function:ProvisionedConcurrency - The provisioned concurrency for a Lambda function.
- cassandra:table:ReadCapacityUnits - The provisioned read capacity for an Amazon Keyspaces table.
- cassandra:table:WriteCapacityUnits - The provisioned write capacity for an Amazon Keyspaces table.
- kafka:broker-storage:VolumeSize - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- elasticache:cache-cluster:Nodes - The number of nodes for an Amazon ElastiCache cache cluster.
- elasticache:replication-group:NodeGroups - The number of node groups for an Amazon ElastiCache replication group.
- elasticache:replication-group:Replicas - The number of replicas per node group for an Amazon ElastiCache replication group.
- neptune:cluster:ReadReplicaCount - The count of read replicas in an Amazon Neptune DB cluster.

- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: Yes

ServiceNamespace

The namespace of the AWS service that provides the resource, or a `custom-resource`.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

PredictedCapacity

The predicted capacity of the scalable target.

Type: Integer

Required: No

ScalableTargetARN

The ARN of the scalable target.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

SuspendedState

Specifies whether the scaling activities for a scalable target are in a suspended state.

Type: SuspendedState 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

ScalableTargetAction

Represents the minimum and maximum capacity for a scheduled action.

Contents

MaxCapacity

The maximum capacity.

Although you can specify a large maximum capacity, note that service quotas may impose lower limits. Each service has its own default quotas for the maximum capacity of the resource. If you want to specify a higher limit, you can request an increase. For more information, consult the documentation for that service. For information about the default quotas for each service, see Service endpoints and quotas in the *Amazon Web Services General Reference*.

Type: Integer

Required: No

MinCapacity

The minimum capacity.

When the scheduled action runs, the resource will have at least this much capacity, but it might have more depending on other settings, such as the target utilization level of a target tracking scaling policy.

For certain resources, the minimum value allowed is 0. For more information, see RegisterScalableTarget.

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

ScalingActivity

Represents a scaling activity.

Contents

ActivityId

The unique identifier of the scaling activity.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

Cause

A simple description of what caused the scaling activity to happen.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

Description

A simple description of what action the scaling activity intends to accomplish.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ResourceId

The identifier of the resource associated with the scaling activity. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.

- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.

- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension

The scalable dimension. This string consists of the service namespace, resource type, and scaling property.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.

- dynamodb:index:WriteCapacityUnits - The provisioned write capacity for a DynamoDB global secondary index.
- rds:cluster:ReadReplicaCount - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- sagemaker:variant:DesiredInstanceCount - The number of EC2 instances for a SageMaker model endpoint variant.
- custom-resource:ResourceType:Property - The scalable dimension for a custom resource provided by your own application or service.
- comprehend:document-classifier-endpoint:DesiredInferenceUnits - The number of inference units for an Amazon Comprehend document classification endpoint.
- comprehend:entity-recognizer-endpoint:DesiredInferenceUnits - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- lambda:function:ProvisionedConcurrency - The provisioned concurrency for a Lambda function.
- cassandra:table:ReadCapacityUnits - The provisioned read capacity for an Amazon Keyspaces table.
- cassandra:table:WriteCapacityUnits - The provisioned write capacity for an Amazon Keyspaces table.
- kafka:broker-storage:VolumeSize - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- elasticache:cache-cluster:Nodes - The number of nodes for an Amazon ElastiCache cache cluster.
- elasticache:replication-group:NodeGroups - The number of node groups for an Amazon ElastiCache replication group.
- elasticache:replication-group:Replicas - The number of replicas per node group for an Amazon ElastiCache replication group.
- neptune:cluster:ReadReplicaCount - The count of read replicas in an Amazon Neptune DB cluster.
- sagemaker:variant:DesiredProvisionedConcurrency - The provisioned concurrency for a SageMaker serverless endpoint.
- sagemaker:inference-component:DesiredCopyCount - The number of copies across an endpoint for a SageMaker inference component.

- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: Yes

ServiceNamespace

The namespace of the AWS service that provides the resource, or a `custom-resource`.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

StartTime

The Unix timestamp for when the scaling activity began.

Type: Timestamp

Required: Yes

StatusCode

Indicates the status of the scaling activity.

Type: String

Valid Values: `Pending | InProgress | Successful | Overridden | Unfulfilled | Failed`

Required: Yes

Details

The details about the scaling activity.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

EndTime

The Unix timestamp for when the scaling activity ended.

Type: Timestamp

Required: No

NotScaledReasons

Machine-readable data that describes the reason for a not scaled activity. Only available when DescribeScalingActivities includes not scaled activities.

Type: Array of NotScaledReason objects

Required: No

StatusMessage

A simple message about the current status of the scaling activity.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

See Also

For more information about using this API in 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

Represents a scaling policy to use with Application Auto Scaling.

For more information about configuring scaling policies for a specific service, see AWS services that you can use with Application Auto Scaling in the *Application Auto Scaling User Guide*.

Contents

CreationTime

The Unix timestamp for when the scaling policy was created.

Type: Timestamp

Required: Yes

PolicyARN

The Amazon Resource Name (ARN) of the scaling policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

PolicyName

The name of the scaling policy.

Type: String

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

Pattern: `\p{Print}+`

Required: Yes

PolicyType

The scaling policy type.

The following policy types are supported:

`TargetTrackingScaling`—Not supported for Amazon EMR

`StepScaling`—Not supported for DynamoDB, Amazon Comprehend, Lambda, Amazon Keyspaces, Amazon MSK, Amazon ElastiCache, or Neptune.

`PredictiveScaling`—Only supported for Amazon ECS

Type: String

Valid Values: `StepScaling | TargetTrackingScaling | PredictiveScaling`

Required: Yes

ResourceId

The identifier of the resource associated with the scaling policy. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.

- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension

The scalable dimension. This string consists of the service namespace, resource type, and scaling property.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.
- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- `sagemaker:variant:DesiredInstanceCount` - The number of EC2 instances for a SageMaker model endpoint variant.
- `custom-resource:ResourceType:Property` - The scalable dimension for a custom resource provided by your own application or service.
- `comprehend:document-classifier-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend document classification endpoint.
- `comprehend:entity-recognizer-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- `lambda:function:ProvisionedConcurrency` - The provisioned concurrency for a Lambda function.
- `cassandra:table:ReadCapacityUnits` - The provisioned read capacity for an Amazon Keyspaces table.

- `cassandra:table:WriteCapacityUnits` - The provisioned write capacity for an Amazon Keyspaces table.
- `kafka:broker-storage:VolumeSize` - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- `elasticache:cache-cluster:Nodes` - The number of nodes for an Amazon ElastiCache cache cluster.
- `elasticache:replication-group:NodeGroups` - The number of node groups for an Amazon ElastiCache replication group.
- `elasticache:replication-group:Replicas` - The number of replicas per node group for an Amazon ElastiCache replication group.
- `neptune:cluster:ReadReplicaCount` - The count of read replicas in an Amazon Neptune DB cluster.
- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency`

`| sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: Yes

ServiceNamespace

The namespace of the AWS service that provides the resource, or a `custom-resource`.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

Alarms

The CloudWatch alarms associated with the scaling policy.

Type: Array of Alarm objects

Required: No

PredictiveScalingPolicyConfiguration

The predictive scaling policy configuration.

Type: PredictiveScalingPolicyConfiguration object

Required: No

StepScalingPolicyConfiguration

A step scaling policy.

Type: StepScalingPolicyConfiguration object

Required: No

TargetTrackingScalingPolicyConfiguration

A target tracking scaling policy.

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

ScheduledAction

Represents a scheduled action.

Contents

CreationTime

The date and time that the scheduled action was created.

Type: Timestamp

Required: Yes

ResourceId

The identifier of the resource associated with the scaling policy. This string consists of the resource type and unique identifier.

- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/my-cluster/my-service`.
- Spot Fleet - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- EMR cluster - The resource type is `instancegroup` and the unique identifier is the cluster ID and instance group ID. Example: `instancegroup/j-2EEZNYKUA1NTV/ig-1791Y4E1L8YI0`.
- AppStream 2.0 fleet - The resource type is `fleet` and the unique identifier is the fleet name. Example: `fleet/sample-fleet`.
- DynamoDB table - The resource type is `table` and the unique identifier is the table name. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the index name. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.
- SageMaker endpoint variant - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.

- Custom resources are not supported with a resource type. This parameter must specify the `OutputValue` from the CloudFormation template stack used to access the resources. The unique identifier is defined by the service provider. More information is available in our GitHub repository.
- Amazon Comprehend document classification endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE`.
- Amazon Comprehend entity recognizer endpoint - The resource type and unique identifier are specified using the endpoint ARN. Example: `arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-endpoint/EXAMPLE`.
- Lambda provisioned concurrency - The resource type is `function` and the unique identifier is the function name with a function version or alias name suffix that is not `$LATEST`. Example: `function:my-function:prod` or `function:my-function:1`.
- Amazon Keyspaces table - The resource type is `table` and the unique identifier is the table name. Example: `keyspace/mykeyspace/table/mytable`.
- Amazon MSK cluster - The resource type and unique identifier are specified using the cluster ARN. Example: `arn:aws:kafka:us-east-1:123456789012:cluster/demo-cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5`.
- Amazon ElastiCache replication group - The resource type is `replication-group` and the unique identifier is the replication group name. Example: `replication-group/mycluster`.
- Amazon ElastiCache cache cluster - The resource type is `cache-cluster` and the unique identifier is the cache cluster name. Example: `cache-cluster/mycluster`.
- Neptune cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:mycluster`.
- SageMaker serverless endpoint - The resource type is `variant` and the unique identifier is the resource ID. Example: `endpoint/my-end-point/variant/KMeansClustering`.
- SageMaker inference component - The resource type is `inference-component` and the unique identifier is the resource ID. Example: `inference-component/my-inference-component`.
- Pool of WorkSpaces - The resource type is `workspacespool` and the unique identifier is the pool ID. Example: `workspacespool/wspool-123456`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

Schedule

The schedule for this action. The following formats are supported:

- At expressions - "`at(`*`yyyy-mm-dd`*`Thh:mm:ss`)"
- Rate expressions - "`rate(`*`value unit`*`)`"
- Cron expressions - "`cron(`*`fields`*`)`"

At expressions are useful for one-time schedules. Cron expressions are useful for scheduled actions that run periodically at a specified date and time, and rate expressions are useful for scheduled actions that run at a regular interval.

At and cron expressions use Universal Coordinated Time (UTC) by default.

The cron format consists of six fields separated by white spaces: [Minutes] [Hours] [Day_of_Month] [Month] [Day_of_Week] [Year].

For rate expressions, *value* is a positive integer and *unit* is `minute` | `minutes` | `hour` | `hours` | `day` | `days`.

For more information, see Schedule recurring scaling actions using cron expressions in the *Application Auto Scaling User Guide*.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScheduledActionARN

The Amazon Resource Name (ARN) of the scheduled action.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScheduledActionName

The name of the scheduled action.

Type: String

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

Pattern: `(?!((^[ ]+.*)|(.*([\u0000-\u001f]|[\u007f-\u009f]|[:/|])+.*)|(.*[ ]+$))).+`

Required: Yes

ServiceNamespace

The namespace of the AWS service that provides the resource, or a `custom-resource`.

Type: String

Valid Values: `ecs | elasticmapreduce | ec2 | appstream | dynamodb | rds | sagemaker | custom-resource | comprehend | lambda | cassandra | kafka | elasticache | neptune | workspaces`

Required: Yes

EndTime

The date and time that the action is scheduled to end, in UTC.

Type: Timestamp

Required: No

ScalableDimension

The scalable dimension. This string consists of the service namespace, resource type, and scaling property.

- `ecs:service:DesiredCount` - The task count of an ECS service.
- `elasticmapreduce:instancegroup:InstanceCount` - The instance count of an EMR Instance Group.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet.

- `appstream:fleet:DesiredCapacity` - The capacity of an AppStream 2.0 fleet.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.
- `sagemaker:variant:DesiredInstanceCount` - The number of EC2 instances for a SageMaker model endpoint variant.
- `custom-resource:ResourceType:Property` - The scalable dimension for a custom resource provided by your own application or service.
- `comprehend:document-classifier-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend document classification endpoint.
- `comprehend:entity-recognizer-endpoint:DesiredInferenceUnits` - The number of inference units for an Amazon Comprehend entity recognizer endpoint.
- `lambda:function:ProvisionedConcurrency` - The provisioned concurrency for a Lambda function.
- `cassandra:table:ReadCapacityUnits` - The provisioned read capacity for an Amazon Keyspaces table.
- `cassandra:table:WriteCapacityUnits` - The provisioned write capacity for an Amazon Keyspaces table.
- `kafka:broker-storage:VolumeSize` - The provisioned volume size (in GiB) for brokers in an Amazon MSK cluster.
- `elasticache:cache-cluster:Nodes` - The number of nodes for an Amazon ElastiCache cache cluster.
- `elasticache:replication-group:NodeGroups` - The number of node groups for an Amazon ElastiCache replication group.
- `elasticache:replication-group:Replicas` - The number of replicas per node group for an Amazon ElastiCache replication group.

- `neptune:cluster:ReadReplicaCount` - The count of read replicas in an Amazon Neptune DB cluster.
- `sagemaker:variant:DesiredProvisionedConcurrency` - The provisioned concurrency for a SageMaker serverless endpoint.
- `sagemaker:inference-component:DesiredCopyCount` - The number of copies across an endpoint for a SageMaker inference component.
- `workspaces:workspacespool:DesiredUserSessions` - The number of user sessions for the WorkSpaces in the pool.

Type: String

Valid Values: `ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | elasticmapreduce:instancegroup:InstanceCount | appstream:fleet:DesiredCapacity | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits | rds:cluster:ReadReplicaCount | sagemaker:variant:DesiredInstanceCount | custom-resource:ResourceType:Property | comprehend:document-classifier-endpoint:DesiredInferenceUnits | comprehend:entity-recognizer-endpoint:DesiredInferenceUnits | lambda:function:ProvisionedConcurrency | cassandra:table:ReadCapacityUnits | cassandra:table:WriteCapacityUnits | kafka:broker-storage:VolumeSize | elasticache:cache-cluster:Nodes | elasticache:replication-group:NodeGroups | elasticache:replication-group:Replicas | neptune:cluster:ReadReplicaCount | sagemaker:variant:DesiredProvisionedConcurrency | sagemaker:inference-component:DesiredCopyCount | workspaces:workspacespool:DesiredUserSessions`

Required: No

ScalableTargetAction

The new minimum and maximum capacity. You can set both values or just one. At the scheduled time, if the current capacity is below the minimum capacity, Application Auto Scaling scales out to the minimum capacity. If the current capacity is above the maximum capacity, Application Auto Scaling scales in to the maximum capacity.

Type: ScalableTargetAction object

Required: No

StartTime

The date and time that the action is scheduled to begin, in UTC.

Type: Timestamp

Required: No

Timezone

The time zone used when referring to the date and time of a scheduled action, when the scheduled action uses an at or cron expression.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

See Also

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

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StepAdjustment

Represents a step adjustment for a StepScalingPolicyConfiguration. Describes an adjustment based on the difference between the value of the aggregated CloudWatch metric and the breach threshold that you've defined for the alarm.

For the following examples, suppose that you have an alarm with a breach threshold of 50:

- To initiate the adjustment when the metric is greater than or equal to 50 and less than 60, specify a lower bound of `0` and an upper bound of `10`.
- To initiate the adjustment when the metric is greater than 40 and less than or equal to 50, specify a lower bound of `-10` and an upper bound of `0`.

There are a few rules for the step adjustments for your step policy:

- The ranges of your step adjustments can't overlap or have a gap.
- At most one step adjustment can have a null lower bound. If one step adjustment has a negative lower bound, then there must be a step adjustment with a null lower bound.
- At most one step adjustment can have a null upper bound. If one step adjustment has a positive upper bound, then there must be a step adjustment with a null upper bound.
- The upper and lower bound can't be null in the same step adjustment.

Contents

ScalingAdjustment

The amount by which to scale, based on the specified adjustment type. A positive value adds to the current capacity while a negative number removes from the current capacity. For exact capacity, you must specify a non-negative value.

Type: Integer

Required: Yes

MetricIntervalLowerBound

The lower bound for the difference between the alarm threshold and the CloudWatch metric. If the metric value is above the breach threshold, the lower bound is inclusive (the metric

must be greater than or equal to the threshold plus the lower bound). Otherwise, it's exclusive (the metric must be greater than the threshold plus the lower bound). A null value indicates negative infinity.

Type: Double

Required: No

MetricIntervalUpperBound

The upper bound for the difference between the alarm threshold and the CloudWatch metric. If the metric value is above the breach threshold, the upper bound is exclusive (the metric must be less than the threshold plus the upper bound). Otherwise, it's inclusive (the metric must be less than or equal to the threshold plus the upper bound). A null value indicates positive infinity.

The upper bound must be greater than the lower bound.

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

StepScalingPolicyConfiguration

Represents a step scaling policy configuration to use with Application Auto Scaling.

For more information, see Step scaling policies in the *Application Auto Scaling User Guide*.

Contents

AdjustmentType

Specifies how the `ScalingAdjustment` value in a StepAdjustment is interpreted (for example, an absolute number or a percentage). The valid values are `ChangeInCapacity`, `ExactCapacity`, and `PercentChangeInCapacity`.

`AdjustmentType` is required if you are adding a new step scaling policy configuration.

Type: String

Valid Values: `ChangeInCapacity | PercentChangeInCapacity | ExactCapacity`

Required: No

Cooldown

The amount of time, in seconds, to wait for a previous scaling activity to take effect. If not specified, the default value is 300. For more information, see Cooldown period in the *Application Auto Scaling User Guide*.

Type: Integer

Required: No

MetricAggregationType

The aggregation type for the CloudWatch metrics. Valid values are `Minimum`, `Maximum`, and `Average`. If the aggregation type is null, the value is treated as `Average`.

Type: String

Valid Values: `Average | Minimum | Maximum`

Required: No

MinAdjustmentMagnitude

The minimum value to scale by when the adjustment type is `PercentChangeInCapacity`. For example, suppose that you create a step scaling policy to scale out an Amazon ECS service by 25 percent and you specify a `MinAdjustmentMagnitude` of 2. If the service has 4 tasks and the scaling policy is performed, 25 percent of 4 is 1. However, because you specified a `MinAdjustmentMagnitude` of 2, Application Auto Scaling scales out the service by 2 tasks.

Type: Integer

Required: No

StepAdjustments

A set of adjustments that enable you to scale based on the size of the alarm breach.

At least one step adjustment is required if you are adding a new step scaling policy configuration.

Type: Array of StepAdjustment 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

SuspendedState

Specifies whether the scaling activities for a scalable target are in a suspended state.

Contents

DynamicScalingInSuspended

Whether scale in by a target tracking scaling policy or a step scaling policy is suspended. Set the value to `true` if you don't want Application Auto Scaling to remove capacity when a scaling policy is triggered. The default is `false`.

Type: Boolean

Required: No

DynamicScalingOutSuspended

Whether scale out by a target tracking scaling policy or a step scaling policy is suspended. Set the value to `true` if you don't want Application Auto Scaling to add capacity when a scaling policy is triggered. The default is `false`.

Type: Boolean

Required: No

ScheduledScalingSuspended

Whether scheduled scaling is suspended. Set the value to `true` if you don't want Application Auto Scaling to add or remove capacity by initiating scheduled actions. The default is `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

TargetTrackingMetric

Represents a specific metric.

Metric is a property of the TargetTrackingMetricStat object.

Contents

Dimensions

The dimensions for the metric. For the list of available dimensions, see the AWS documentation available from the table in AWS services that publish CloudWatch metrics in the *Amazon CloudWatch User Guide*.

Conditional: If you published your metric with dimensions, you must specify the same dimensions in your scaling policy.

Type: Array of TargetTrackingMetricDimension objects

Required: No

MetricName

The name of the metric.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

Namespace

The namespace of the metric. For more information, see the table in AWS services that publish CloudWatch metrics in the *Amazon CloudWatch User Guide*.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

See Also

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

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TargetTrackingMetricDataQuery

The metric data to return. Also defines whether this call is returning data for one metric only, or whether it is performing a math expression on the values of returned metric statistics to create a new time series. A time series is a series of data points, each of which is associated with a timestamp.

You can use `TargetTrackingMetricDataQuery` structures with a PutScalingPolicy operation when you specify a TargetTrackingScalingPolicyConfiguration in the request.

You can call for a single metric or perform math expressions on multiple metrics. Any expressions used in a metric specification must eventually return a single time series.

For more information and examples, see Create a target tracking scaling policy for Application Auto Scaling using metric math in the *Application Auto Scaling User Guide*.

Contents

Id

A short name that identifies the object's results in the response. This name must be unique among all `MetricDataQuery` objects specified for a single scaling policy. If you are performing math expressions on this set of data, this name represents that data and can serve as a variable in the mathematical expression. The valid characters are letters, numbers, and underscores. The first character must be a lowercase letter.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

Expression

The math expression to perform on the returned data, if this object is performing a math expression. This expression can use the `Id` of the other metrics to refer to those metrics, and can also use the `Id` of other expressions to use the result of those expressions.

Conditional: Within each `TargetTrackingMetricDataQuery` object, you must specify either `Expression` or `MetricStat`, but not both.

Type: String

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

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

Label

A human-readable label for this metric or expression. This is especially useful if this is a math expression, so that you know what the value represents.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

MetricStat

Information about the metric data to return.

Conditional: Within each `MetricDataQuery` object, you must specify either `Expression` or `MetricStat`, but not both.

Type: TargetTrackingMetricStat object

Required: No

ReturnData

Indicates whether to return the timestamps and raw data values of this metric.

If you use any math expressions, specify `true` for this value for only the final math expression that the metric specification is based on. You must specify `false` for `ReturnData` for all the other metrics and expressions used in the metric specification.

If you are only retrieving metrics and not performing any math expressions, do not specify anything for `ReturnData`. This sets it to its default (`true`).

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

TargetTrackingMetricDimension

Describes the dimension of a metric.

Contents

Name

The name of the dimension.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

Value

The value of the dimension.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\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

TargetTrackingMetricStat

This structure defines the CloudWatch metric to return, along with the statistic and unit.

`TargetTrackingMetricStat` is a property of the TargetTrackingMetricDataQuery object.

For more information about the CloudWatch terminology below, see Amazon CloudWatch concepts in the *Amazon CloudWatch User Guide*.

Contents

Metric

The CloudWatch metric to return, including the metric name, namespace, and dimensions. To get the exact metric name, namespace, and dimensions, inspect the Metric object that is returned by a call to ListMetrics.

Type: TargetTrackingMetric object

Required: Yes

Stat

The statistic to return. It can include any CloudWatch statistic or extended statistic. For a list of valid values, see the table in Statistics in the *Amazon CloudWatch User Guide*.

The most commonly used metric for scaling is `Average`.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

Unit

The unit to use for the returned data points. For a complete list of the units that CloudWatch supports, see the MetricDatum data type in the *Amazon CloudWatch API Reference*.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1023.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

See Also

For more information about using this API in 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

Represents a target tracking scaling policy configuration to use with Application Auto Scaling.

For more information, see Target tracking scaling policies in the *Application Auto Scaling User Guide*.

Contents

TargetValue

The target value for the metric. Although this property accepts numbers of type Double, it won't accept values that are either too small or too large. Values must be in the range of -2^360 to 2^360. The value must be a valid number based on the choice of metric. For example, if the metric is CPU utilization, then the target value is a percent value that represents how much of the CPU can be used before scaling out.

> **Note**
>
> If the scaling policy specifies the `ALBRequestCountPerTarget` predefined metric, specify the target utilization as the optimal average request count per target during any one-minute interval.

Type: Double

Required: Yes

CustomizedMetricSpecification

A customized metric. You can specify either a predefined metric or a customized metric.

Type: CustomizedMetricSpecification object

Required: No

DisableScaleIn

Indicates whether scale in by the target tracking scaling policy is disabled. If the value is `true`, scale in is disabled and the target tracking scaling policy won't remove capacity from the scalable target. Otherwise, scale in is enabled and the target tracking scaling policy can remove capacity from the scalable target. The default value is `false`.

Type: Boolean

Required: No

PredefinedMetricSpecification

A predefined metric. You can specify either a predefined metric or a customized metric.

Type: PredefinedMetricSpecification object

Required: No

ScaleInCooldown

The amount of time, in seconds, after a scale-in activity completes before another scale-in activity can start. For more information and for default values, see Define cooldown periods in the *Application Auto Scaling User Guide*.

Type: Integer

Required: No

ScaleOutCooldown

The amount of time, in seconds, to wait for a previous scale-out activity to take effect. For more information and for default values, see Define cooldown periods in the *Application Auto Scaling User Guide*.

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

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