Manual do usuário

Amazon EC2 Auto Scaling

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

Amazon EC2 Auto Scaling: Manual do usuário

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

As marcas comerciais e imagens comerciais da Amazon não podem ser usadas no contexto de nenhum produto ou serviço que não seja da Amazon, nem de qualquer maneira que possa gerar confusão entre os clientes ou que deprecie ou desprestigie a Amazon. Todas as outras marcas comerciais que não pertencem à Amazon pertencem a seus respectivos proprietários, que podem ou não ser afiliados, patrocinados pela Amazon ou ter conexão com ela.

Table of Contents

O que é o Amazon EC2 Auto Scaling? 1
Atributos do Amazon EC2 Auto Scaling 1
Preços do Amazon EC2 Auto Scaling 3
Conceitos básicos 4
Trabalhar com grupos do Auto Scaling 4
Benefícios do Auto Scaling 5
Exemplo: atender a demanda variável 5
Exemplo: arquitetura de aplicação Web 7
Exemplo: distribuir instâncias entre zonas de disponibilidade 9
Ciclo de vida da instância 12
Escalonamento horizontal 13
Instâncias em serviço 13
Reduzir a escala na horizontal 14
Desvincular uma instância 15
Anexar uma instância 15
Hooks do ciclo de vida 15
Entrar e sair de espera 16
Cotas do Amazon EC2 Auto Scaling 16
Limitação de controle de utilização para a API do Amazon EC2 Auto Scaling 18
Taxas de encerramento do EC2 19
Outros produtos da 19
Configurar 20
Prepare-se para usar o AWS CLI 20
Conceitos básicos 21
Tutorial: Criar seu primeiro grupo do Auto Scaling 22
Preparar para a demonstração 22
Etapa 1: Criar um modelo de execução 23
Etapa 2: Criar um grupo do Auto Scaling com uma única instância 24
Etapa 3: Verificar seu grupo do Auto Scaling 25
Etapa 4: Terminar uma instância no seu grupo do Auto Scaling 26
Etapa 5: Próximas etapas 27
Etapa 6: limpar 27
Tutorial: Configurar uma aplicação escalonada e com balanceamento de carga 28
Pré-requisitos 30

Etapa 1: Configurar um modelo de execução ou uma configuração de execução 31
Etapa 2: Criar um grupo do Auto Scaling 35
Etapa 3: Verificar se o balanceador de carga está anexado 36
Etapa 4: Próximas etapas 37
Etapa 5: Limpar 38
Recursos relacionados 39
Modelos de inicialização 40
Permissões para trabalhar com modelos de inicialização 41
Operações de API compatíveis com os modelos de execução 41
Criar um modelo de execução para um grupo do Auto Scaling 41
Criar seu modelo de execução (console) 42
Alterar as configurações da interface de rede padrão (console) 45
Modificar a configuração do armazenamento (console) 47
Criar um modelo de execução com base em uma instância existente (console) 51
Recursos relacionados 51
Limitações 51
Criar um modelo de execução usando configurações avançadas 52
Configurações necessárias 52
Configurações avançadas 53
Request Spot Instances 57
Capacity Blocks para ML 59
Migre seus grupos do Auto Scaling para modelos de execução 65
Etapa 1: encontrar grupos do Auto Scaling que usem configurações de execução 65
Etapa 2: copiar uma configuração de execução para um modelo de execução 68
Etapa 3: atualizar um grupo do Auto Scaling para usar um modelo de execução 69
Etapa 4: substituir suas instâncias 70
Mais informações 70
Migre CloudFormation pilhas para modelos de lançamento 70
Encontre grupos do Auto Scaling que usam uma configuração de execução 71
Atualizar uma pilha para usar um modelo de execução 72
Compreender atualização de comportamentos de recursos da pilha 76
Rastrear a migração 77
Referência do mapeamento de configuração de execução 77
AWS CLI exemplos para trabalhar com modelos de lançamento 79
Exemplo de uso 80
Criar um modelo de execução básico 80

Especificar etiquetas que marcam instâncias ao iniciar ... 81
Especificar um perfil do IAM a ser transmitido às instâncias ... 82
Atribuir um endereço IP público ... 82
Especificar um script de dados do usuário que configura instâncias ao iniciar ... 83
Especificar um mapeamento de dispositivos de blocos ... 83
Especificar hosts dedicados para trazer licenças de software de fornecedores externos ... 83
Especificar uma interface de rede existente ... 84
Criar várias interfaces de rede ... 84
Gerenciar modelos de execução ... 85
Atualizar um grupo do Auto Scaling para usar um modelo de execução ... 88
Use os parâmetros do Systems Manager em vez da AMI IDs ... 89
Criar um modelo de execução que especifique um parâmetro para a AMI ... 89
Verifique se um modelo de execução obtém o ID da AMI correto ... 94
Recursos relacionados ... 95
Limitações ... 95
Configurações de lançamento ... 97
Criar uma configuração de execução ... 98
Criar uma configuração de execução ... 99
Configurar IMDS ... 102
Criar uma configuração de execução usando uma instância do EC2 ... 104
Alterar uma configuração de execução ... 109
Grupos do Auto Scaling ... 111
Criar grupos do Auto Scaling usando modelos de execução ... 113
Criar um grupo usando um modelo de execução ... 113
Criar um grupo usando o assistente de execução do EC2 ... 117
Usar vários tipos de instâncias e opções de compra ... 122
Criar grupos do Auto Scaling usando configurações de execução ... 172
Criar um grupo usando uma configuração de execução ... 173
Crie um grupo a partir da instância usando AWS CLI ... 176
Provisionamento síncrono ... 181
Provisionamento síncrono e escalabilidade assíncrona ... 182
Limitações e considerações ... 183
Lançamento de instâncias com provisionamento síncrono ... 185
Atualizar um grupo do Auto Scaling ... 188
Atualizar instâncias do Auto Scaling ... 189
Estratégia de alocação de grupos do Auto Scaling e alterações de capacidade ... 190

Marcar grupos e instâncias ... 191
Restrições de nomeação e uso de tags ... 192
Ciclo de vida de marcação de instâncias do EC2 ... 193
Marcar seus grupos do Auto Scaling ... 193
Excluir tags ... 196
Etiquetas para segurança ... 197
Controlar o acesso usando etiquetas ... 198
Usar etiquetas para filtrar grupos do Auto Scaling ... 199
Políticas de manutenção de instância ... 203
Visão geral do ... 204
Defina uma política de manutenção de instância no seu grupo ... 212
Ganchos do ciclo de vida ... 216
Disponibilidade de ganchos do ciclo de vida ... 218
Considerações e limitações ... 218
Recursos relacionados ... 221
Como os ganchos do ciclo de vida funcionam em grupos do Auto Scaling ... 221
Preparar para adicionar um gancho do ciclo de vida ... 225
Controle a retenção de instâncias com políticas de ciclo de vida de instâncias ... 235
Recuperar o estado de destino do ciclo de vida ... 240
Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling ... 242
Concluir uma ação do ciclo de vida em um grupo do Auto Scaling ... 246
Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância ... 248
Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda ... 258
Pools ativos ... 268
Principais conceitos ... 268
Pré-requisitos ... 271
Atualização das instâncias em um pool ativo ... 272
Recursos relacionados ... 273
Limitações ... 273
Usar ganchos do ciclo de vida ... 274
Criar um pool ativo para seu grupo do Auto Scaling ... 279
Visualizar status da verificação de integridade ... 281
AWS CLI exemplos para trabalhar com piscinas aquecidas ... 285
Deslocamento zonal do grupo Auto Scaling ... 287
Conceitos de mudança zonal de grupos do Auto Scaling ... 288

Como a mudança zonal funciona para grupos de Auto Scaling ... 289
Melhores práticas para usar o deslocamento zonal ... 291
Ative o deslocamento zonal usando o Console de gerenciamento da AWS ou o AWS CLI ... 292
Distribuição de zonas de disponibilidade ... 295
Instâncias de desanexação e anexação ... 295
Considerações para desanexar instâncias ... 296
Considerações para anexar instâncias ... 297
Mover uma instância para um grupo diferente usando desanexar e anexar ... 297
Remover instâncias temporariamente ... 302
Como o estado de espera funciona ... 303
Considerações ... 304
Status de integridade de uma instância em um estado de espera ... 305
Remover temporariamente uma instância definindo-a como em espera ... 303
Excluir infraestrutura do Auto Scaling ... 309
Excluir seu grupo do Auto Scaling ... 310
(Opcional) Excluir a configuração de execução ... 311
(Opcional) Excluir o modelo de execução ... 311
(Opcional) Excluir o balanceador de carga e grupos de destino ... 312
(Opcional) Excluir CloudWatch alarmes ... 313
Proteção contra exclusão de recursos ... 314
Substitua suas instâncias ... 321
Atualização de instância ... 322
Como funciona a atualização de uma instância ... 323
Entender os valores padrão ... 330
Iniciar uma atualização de instância ... 333
Monitore a atualização de uma instância ... 346
Substitua os volumes raiz ... 349
Cancelar uma atualização de instância ... 354
Desfazer alterações com uma reversão ... 355
Usar a opção de ignorar correspondência ... 360
Adicionar pontos de verificação ... 370
Vida útil máxima da instância ... 376
Considerações ... 376
Definir o tempo de vida máximo da instância ... 377
Limitações ... 378
Escalar o grupo ... 380

Escolha seu método de escalabilidade 381
Definir limites de escalabilidade 382
Definir o aquecimento de instância padrão 384
Considerações sobre o desempenho de escalabilidade 385
Escolha o tempo para o aquecimento de instância padrão 386
Habilitar o aquecimento de instância padrão para um grupo 387
Verificar o aquecimento de instância padrão para um grupo 389
Encontre políticas de escalabilidade com um tempo de aquecimento da instância definido anteriormente 390
Limpe o aquecimento da instância definido anteriormente para uma política de escalabilidade 391
Escalabilidade manual 391
Alterar a capacidade desejada de seu grupo do Auto Scaling 392
Encerrar uma instância no seu grupo do Auto Scaling (AWS CLI) 396
Escalabilidade programada 397
Como a escalabilidade programada funciona 398
Programações recorrentes 398
Fuso horário 399
Considerações 400
Limitações 401
Criar uma ação programada 401
Visualizar detalhes da ação programada 403
Excluir uma ação programada 404
Escalabilidade dinâmica 405
Como funcionam as políticas de escalabilidade dinâmica 406
Várias políticas de escalabilidade dinâmica 407
Políticas de escalabilidade de rastreamento de destino 409
Políticas de escalabilidade simples e em etapas 427
Desaquecimento de escalabilidade 445
Ajuste de escala baseado no Amazon SQS 449
Verificar uma ação de escalabilidade 457
Desabilitar uma política de escalabilidade 459
Desabilitar uma política de escalabilidade para um grupo do Auto Scaling 462
AWS CLI exemplos de políticas de escalabilidade 464
Escalabilidade preditiva 467
Como a escalabilidade preditiva funciona 468

Criar uma política de escalabilidade preditiva () ... 473
Avaliar as políticas de escalabilidade preditiva ... 481
Substituir a previsão ... 490
Usar métricas personalizadas ... 496
Controlar o término de instâncias ... 508
Cenários de término ... 508
Configure políticas de terminação ... 512
Criar uma política de término personalizada com o Lambda ... 518
Usar proteção de redução na escala na horizontal de instâncias ... 526
Criar para finalização de instância sem problemas ... 530
Suspender-retomar processos ... 539
Tipos de processos ... 539
Considerações ... 541
Suspender processos ... 541
Retomar processos ... 542
Entender como a suspensão de processos afeta outros processos ... 543
Monitorar ... 548
Verificações de integridade ... 550
Sobre verificações de integridade ... 551
Defina o período de carência da verificação de integridade ... 558
Monitoramento de volumes do Amazon EBS com problemas ... 561
Configuração de uma verificação de integridade personalizada ... 565
Veja o motivo das falhas na verificação de integridade ... 566
Solução de problemas de instâncias não íntegras ... 568
Monitor com Health Dashboard ... 572
Monitore CloudWatch métricas ... 573
Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling ... 574
CloudWatch métricas para Amazon EC2 Auto Scaling ... 579
Configurar monitoramento para instâncias do Auto Scaling ... 587
Registre chamadas de API usando CloudTrail ... 590
Eventos de gerenciamento do Auto Scaling em CloudTrail ... 591
Exemplos do Auto Scaling usando o ... 591
Chamadas de Auto Scaling ativadas RemoveAction CloudWatch ... 593
Notificações do Amazon SNS ... 593
Amazon SNS e Amazon EC2 Auto Scaling ... 593
Trabalhar com outros serviços ... 600

Rebalanceamento de capacidade 600
Visão geral do 601
Comportamento de rebalanceamento de capacidade 602
Considerações 603
Habilitar o rebalanceamento de capacidade 605
Reservas de capacidade 611
Preferência de reserva de capacidade 612
Saldo da zona de disponibilidade e reservas de capacidade 613
Uso da preferência de reserva de capacidade 614
Reservas de capacidade 616
AWS CloudShell 621
CloudFormation 621
Amazon EC2 CloudFormation Auto Scaling e modelos 622
Saiba mais sobre CloudFormation 622
Compute Optimizer 623
Limitações 623
Conclusões 624
Exibir recomendações 624
Considerações para avaliação das recomendações 625
Elastic Load Balancing 626
Tipos de Elastic Load Balancing 627
Preparar para anexar um balanceador de carga 629
Anexar um balanceador de carga 632
Configurar o balanceador de carga 636
Verificar o status do anexo 637
Adicione uma Zona de disponibilidade 638
Remover uma Zona de disponibilidade 641
Desvincular um balanceador de carga 635
AWS CLI exemplos para trabalhar com o Elastic Load Balancing 643
Estrutura de PVC 650
Preparar para anexar um grupo de destino 653
Anexar um grupo de destino do VPC Lattice 656
Verificar o status do anexo 661
EventBridge 662
Referência de eventos do Amazon EC2 Auto Scaling 663
Exemplos de eventos e padrões de atualização de instâncias 669

Exemplos de eventos e padrões de pool ativo ... 677
EventBridge regras ... 683
Amazon VPC ... 688
VPC padrão ... 689
VPC não padrão ... 689
Considerações sobre a escolha de sub-redes da VPC ... 689
Endereçamento IP em uma VPC ... 690
Interfaces de rede em uma VPC ... 691
Locação de localização de instância ... 691
AWS Outposts ... 691
Mais recursos para aprender sobre VPCs ... 692
Segurança ... 693
Segurança da infraestrutura ... 694
Recursos relacionados ... 694
Resiliência ... 694
Recursos relacionados ... 696
Proteção de dados ... 696
Use AWS KMS keys para criptografar volumes do Amazon EBS ... 697
Recursos relacionados ... 698
AWS KMS política chave para uso com volumes criptografados ... 698
Gerenciamento de Identidade e Acesso ... 705
Controle de acesso ... 705
Como o Amazon EC2 Auto Scaling funciona com o IAM ... 706
Permissões de API ... 714
Políticas gerenciadas ... 716
Perfis vinculados ao serviço ... 722
Exemplos de políticas baseadas em identidade ... 727
Prevenção do problema "confused deputy" entre serviços ... 737
Controle o uso do modelo de execução do Amazon EC2 em grupos do Auto Scaling ... 740
Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2 ... 749
Validação de conformidade ... 752
Conformidade do PCI DSS ... 752
Usar endpoints da VPC para conectividade privada ... 752
Criar um VPC endpoint de interface ... 753
Criar uma política de endpoint da VPC ... 753
Trabalhando com AWS SDKs ... 755

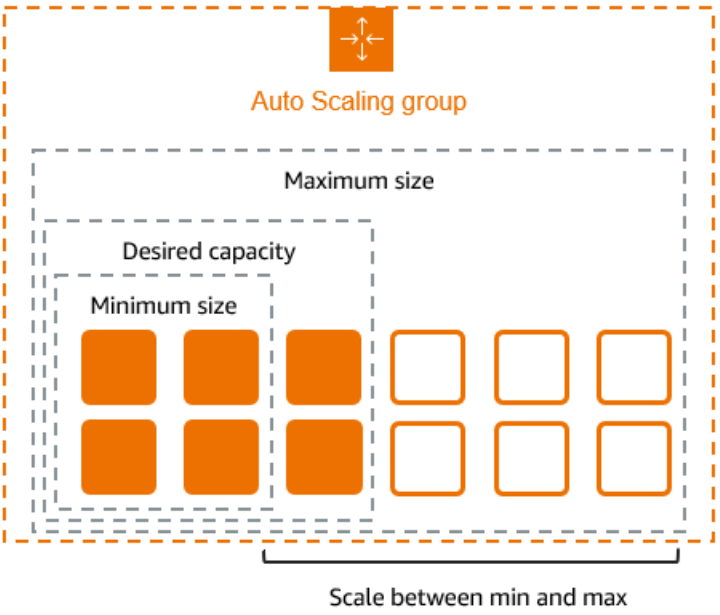
Trabalhando com as APIs 757
Exemplos de código 761
Conceitos básicos 763
Olá, Auto Scaling 765
Conheça os conceitos básicos 775
Ações 877
Cenários 1110
Criar e gerenciar um serviço resiliente 1110
Solução de problemas 1279
Recuperar uma mensagem de erro 1279
Desativar atividades de escala 1281
Recursos adicionais para solução de problemas 1282
Falha ao iniciar instância 1283
A configuração solicitada não é suportada atualmente. 1284
O grupo de segurança <nome do grupo de segurança> não existe. Falha ao ativar a instância do EC2. 1285
O par de chaves <par de chaves associado à sua instância do EC2> não existe. Falha ao ativar a instância do EC2. 1285
O tipo de instância solicitado (<tipo de instância>) não tem suporte na Zona de disponibilidade solicitada (<Zona de disponibilidade da instância>)... 1286
Seu preço de solicitação spot de 0,015 é inferior ao preço mínimo de atendimento de solicitação spot exigido de 0,0735... 1286
Nome de dispositivo inválido <nome do dispositivo> / Carregamento do nome de dispositivo inválido. Falha ao ativar a instância do EC2. 1286
O valor (<nome associado ao dispositivo de armazenamento de instâncias>) do parâmetro virtualName é inválido... Falha ao ativar a instância do EC2. 1287
Os mapeamentos de dispositivos de blocos do EBS não são compatíveis com o armazenamento de instâncias. AMIs 1288
Os grupos de posicionamento não podem ser usados com instâncias do tipo '<instance type>'. Falha ao ativar a instância do EC2. 1288
Cliente. InternalError: Erro do cliente na inicialização. 1288
No momento, não temos capacidade de <tipo de instância> suficiente para tipo de instância na zona de disponibilidade solicitada. Falha ao ativar a instância do EC2. 1290
A reserva solicitada não tem capacidade compatível e disponível suficiente para essa solicitação. Falha ao ativar a instância do EC2. 1290

Sua reserva do bloco de capacidade <reservation id> ainda não está ativa. Falha ao ativar a instância do EC2. 1291
Não há capacidade spot disponível que corresponda à sua solicitação. Falha ao ativar a instância do EC2. 1292
<número de instâncias> instância(s) já estão em execução. Falha ao ativar a instância do EC2. 1292
Problemas de AMI 1293
O ID da AMI <ID de sua AMI> não existe. Falha ao ativar a instância do EC2. 1293
A AMI <ID da AMI> está pendente e não pode ser executada. Falha ao ativar a instância do EC2. 1294
Nome do dispositivo inválido <device name>. Falha ao ativar a instância do EC2. 1294
A arquitetura 'arm64' do tipo de instância especificado não corresponde à arquitetura 'x86_64' da AMI especificada... Falha na execução da instância do EC2. 1295
A AMI '<AMI ID>' está desabilitada e não pode ser executada. Falha ao ativar a instância do EC2. 1296
Problemas do balanceador de carga 1297
Um ou mais grupos de destino não encontrados. Falha na validação da configuração do balanceador de carga. 1298
Não é possível encontrar o Load Balancer <seu load balancer>. Falha na validação da configuração do balanceador de carga. 1298
Não há nenhum balanceador de carga ATIVO chamado <nome do balanceador de carga>. Falha ao atualizar a configuração do balanceador de carga. 1299
A instância do EC2 <ID da instância> não está na VPC. Falha ao atualizar a configuração do balanceador de carga. 1299
Problemas em modelos de execução 1299
É necessário usar um modelo de inicialização totalmente formado válido (valor inválido) .. 1299
Você não está autorizado a usar o modelo de execução (permissões insuficientes) 1300
Informações relacionadas 1302
Histórico do documento 1304
..... mcccl

O que é o Amazon EC2 Auto Scaling?

O Amazon EC2 Auto Scaling ajuda a garantir que você tenha o número correto de instâncias do Amazon EC2 disponíveis para processar a carga da sua aplicação. Você cria coleções de instâncias do EC2, chamadas de grupos de Auto Scaling. Você pode especificar o número mínimo de instâncias em cada grupo do Auto Scaling, e o Amazon EC2 Auto Scaling garante que seu grupo nunca seja menor que esse tamanho. Você pode especificar o número máximo de instâncias em cada grupo do Auto Scaling, e o Amazon EC2 Auto Scaling garante que seu grupo nunca seja maior que esse tamanho. Se você especificar a capacidade desejada, quando você criar o grupo ou em qualquer momento depois disso, o Amazon EC2 Auto Scaling garante que seu grupo tenha essa quantidade de instâncias. Se você especificar políticas de escalabilidade, o Amazon EC2 Auto Scaling poderá iniciar ou terminar instâncias à medida que a demanda da aplicação aumentar ou diminuir.

Por exemplo, o seguinte grupo do Auto Scaling tem um tamanho mínimo de quatro instâncias, uma capacidade desejada de seis instâncias e um tamanho máximo de doze instâncias. As políticas de escalabilidade que você define ajustam o número de instâncias, em seu número mínimo e máximo de instâncias, com base nos critérios que você especifica.



Atributos do Amazon EC2 Auto Scaling

Com o Amazon EC2 Auto Scaling, suas instâncias do EC2 são organizadas em grupos de Auto Scaling para que possam ser tratadas como uma unidade lógica para fins de escalabilidade e

gerenciamento. Los grupos de Auto Scaling utilizan plantillas de lanzamiento (o configuraciones de lanzamiento) como plantillas de configuración para sus instancias EC2.

Estes são os principais atributos do Amazon EC2 Auto Scaling:

Monitoramento da integridade das instâncias em execução

O Amazon EC2 Auto Scaling monitora automaticamente a integridade e a disponibilidade de suas instâncias usando as verificações de integridade do EC2 e substitui as instâncias terminadas ou com problemas para manter a capacidade desejada.

Verificações de integridade personalizadas

Além das verificações de integridade incorporadas, você pode definir verificações de integridade personalizadas específicas da sua aplicação para verificar se está respondendo conforme o esperado. Se uma instância falhar em sua verificação de integridade personalizada, ela será automaticamente substituída para manter a capacidade desejada.

Equilíbrio entre Zonas de disponibilidade.

Você pode especificar várias zonas de disponibilidade para o seu grupo do Auto Scaling, e o Amazon EC2 Auto Scaling equilibra suas instâncias uniformemente entre as zonas de disponibilidade à medida que o grupo escala. Isso proporciona alta disponibilidade e resiliência, protegendo suas aplicações contra falhas em um único local.

Vários tipos de instâncias e várias opções de compra

Em um único grupo do Auto Scaling, você pode lançar vários tipos de instâncias e opções de compra (instâncias spot e sob demanda), permitindo otimizar os custos com o uso de instâncias spot. Você também pode aproveitar os descontos da instância reservada e do Savings Plans usando-os com as instâncias sob demanda no grupo.

Substituição automatizada de instâncias spot

Se o seu grupo incluir instâncias spot Amazon EC2 Auto Scaling, ela poderá solicitar automaticamente a reposição da capacidade spot se as instâncias spot forem interrompidas. Por meio do rebalanceamento da capacidade, o Amazon EC2 Auto Scaling também pode monitorar e substituir proativamente suas instâncias spot que estão em alto risco de interrupção.

Balanceamento de carga

Você pode usar o balanceamento de carga e as verificações de integridade do Elastic Load Balancing para garantir uma distribuição uniforme do tráfego de aplicações para suas instâncias

íntegras. Sempre que as instâncias são iniciadas ou terminadas, o Amazon EC2 Auto Scaling registra e cancela automaticamente o registro das instâncias no balanceador de carga.

Escalabilidade

O Amazon EC2 Auto Scaling também fornece várias formas de escalar seus grupos do Auto Scaling. O uso do ajuste de escala automático permite que você mantenha a disponibilidade da aplicação e reduza os custos, adicionando capacidade para lidar com cargas de pico e removendo a capacidade quando a demanda for menor. Você também pode ajustar manualmente o tamanho do grupo do Auto Scaling conforme necessário.

Atualização de instância

O atributo de atualização de instância fornece um mecanismo para atualizar as instâncias de forma contínua quando você atualiza a AMI ou o modelo de execução. Você também pode usar uma abordagem em fases, conhecida como implantação canário, para testar uma nova AMI ou modelo de execução em um pequeno conjunto de instâncias antes de implementá-la em todo o grupo.

Hooks do ciclo de vida

Os hooks do ciclo de vida são úteis para definir ações personalizadas que são chamadas quando novas instâncias são iniciadas ou antes que as instâncias sejam terminadas. Esse atributo é particularmente útil para a criação de arquiteturas orientadas por eventos, mas também ajuda a gerenciar instâncias ao longo de seu ciclo de vida.

Suporte para cargas de trabalho com estado

Os hooks de ciclo de vida também oferecem um mecanismo para manter o estado no desligamento. Para garantir a continuidade das aplicações com estado, você também pode usar a proteção de redução da escala horizontalmente ou políticas de terminação personalizadas para evitar que instâncias com processos de longa duração sejam terminadas antecipadamente.

Para obter mais informações sobre os benefícios do Amazon EC2 Auto Scaling consulte Benefícios do Auto Scaling para a arquitetura de aplicações.

Preços do Amazon EC2 Auto Scaling

Não há taxas adicionais com o Amazon EC2 Auto Scaling, então é fácil testá-lo e ver como ele pode AWS beneficiar sua arquitetura. Você paga somente pelos AWS recursos (por exemplo, instâncias do EC2, volumes do EBS e CloudWatch alarmes) que você usa.

Conceitos básicos

Para começar, complete o tutorial Criar seu primeiro grupo de Auto Scaling para criar um grupo de Auto Scaling e ver como ele responde quando uma instância nesse grupo é encerrada.

Trabalhar com grupos do Auto Scaling

Você pode criar, acessar e gerenciar seus grupos do Auto Scaling usando qualquer uma das seguintes interfaces:

- Console de gerenciamento da AWS – fornece uma interface da Web que você pode usar para acessar os grupos do Auto Scaling. Se você se inscreveu em um Conta da AWS, você pode acessar seus grupos de Auto Scaling fazendo login no Console de gerenciamento da AWS, usando a caixa de pesquisa na barra de navegação para pesquisar grupos de Auto Scaling e, em seguida, escolhendo grupos de Auto Scaling.
- AWS Command Line Interface (AWS CLI) — Fornece comandos para um amplo conjunto de Serviços da AWS e é compatível com Windows, macOS e Linux. Para começar, consulte o Prepare-se para usar o AWS CLI. Para obter mais informações, consulte escalabilidade automática na Referência de comandos da AWS CLI .
- AWS Tools for Windows PowerShell— Fornece comandos para um amplo conjunto de AWS produtos para quem cria scripts no PowerShell ambiente. Para começar a usar, consulte o Guia do usuário do Ferramentas da AWS para PowerShell. Para obter mais informações, consulte Referência de Cmdlets do Ferramentas da AWS para PowerShell.
- AWS SDKs— fornece operações de API específicas do idioma e cuida de muitos detalhes da conexão, como calcular assinaturas, lidar com novas tentativas de solicitação e lidar com erros. Para obter mais informações, consulte AWS SDKs.
- API de consulta: fornece ações de API de baixo nível que são chamadas usando solicitações HTTPS. Usar a API de consulta é a maneira mais direta de acessar a Serviços da AWS. No entanto, ela exige que a aplicação trate detalhes de baixo nível, como gerar o hash para assinar a solicitação e tratar erros. Para obter mais informações, consulte a Referência da API do Amazon EC2 Auto Scaling.
- CloudFormation— Suporta a criação de grupos de Auto Scaling usando CloudFormation modelos. Para obter mais informações, consulte Crie grupos de Auto Scaling com AWS CloudFormation.

Para se conectar programaticamente a um AWS service (Serviço da AWS), você usa um endpoint. .

Benefícios do Auto Scaling para a arquitetura de aplicações

Adicionar o Amazon EC2 Auto Scaling à sua arquitetura de aplicativos é uma forma de maximizar os benefícios AWS da nuvem. Quando o Amazon EC2 Auto Scaling é usado, suas aplicações obtêm os seguintes benefícios:

- Melhor tolerância a falhas. O Amazon EC2 Auto Scaling pode detectar quando uma instância não está íntegra, terminá-la e iniciar uma instância para substituí-la. Você também pode configurar o Amazon EC2 Auto Scaling para usar várias zonas de disponibilidade. Se uma zona de disponibilidade se tornar indisponível, o Amazon EC2 Auto Scaling poderá iniciar instâncias em outra zona para compensar.
- Melhor disponibilidade. O Amazon EC2 Auto Scaling ajuda a garantir que a aplicação sempre tenha a capacidade certa para lidar com a demanda de tráfego atual.
- Melhor gerenciamento de custos. O Amazon EC2 Auto Scaling pode aumentar e reduzir dinamicamente a capacidade, conforme necessário. Como você paga pelas instâncias do EC2 que usa, você pode economizar ativando instâncias quando elas são realmente necessárias e encerrando-as quando não são necessárias.

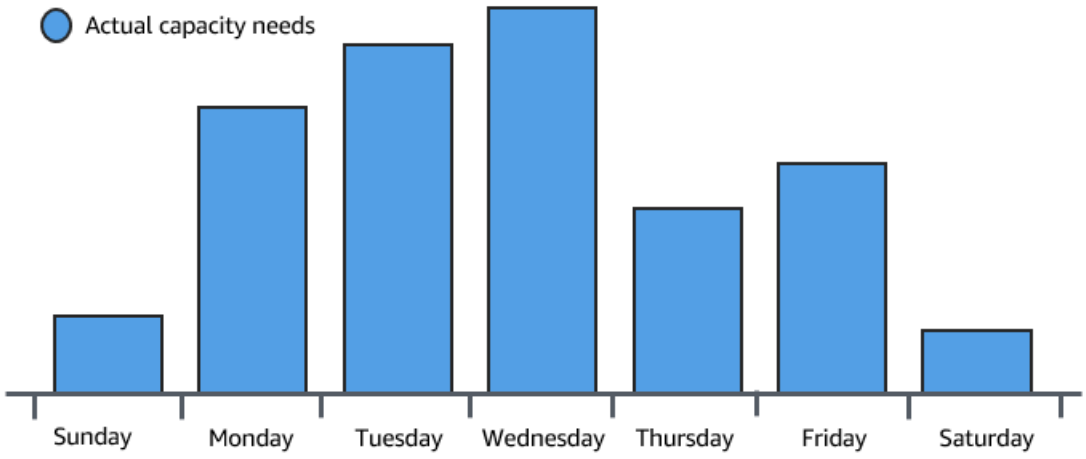
Conteúdo

- Exemplo: atender a demanda variável
- Exemplo: arquitetura de aplicação Web
- Exemplo: distribuir instâncias entre zonas de disponibilidade
  - Distribuição de instâncias
  - Atividades de rebalanceamento

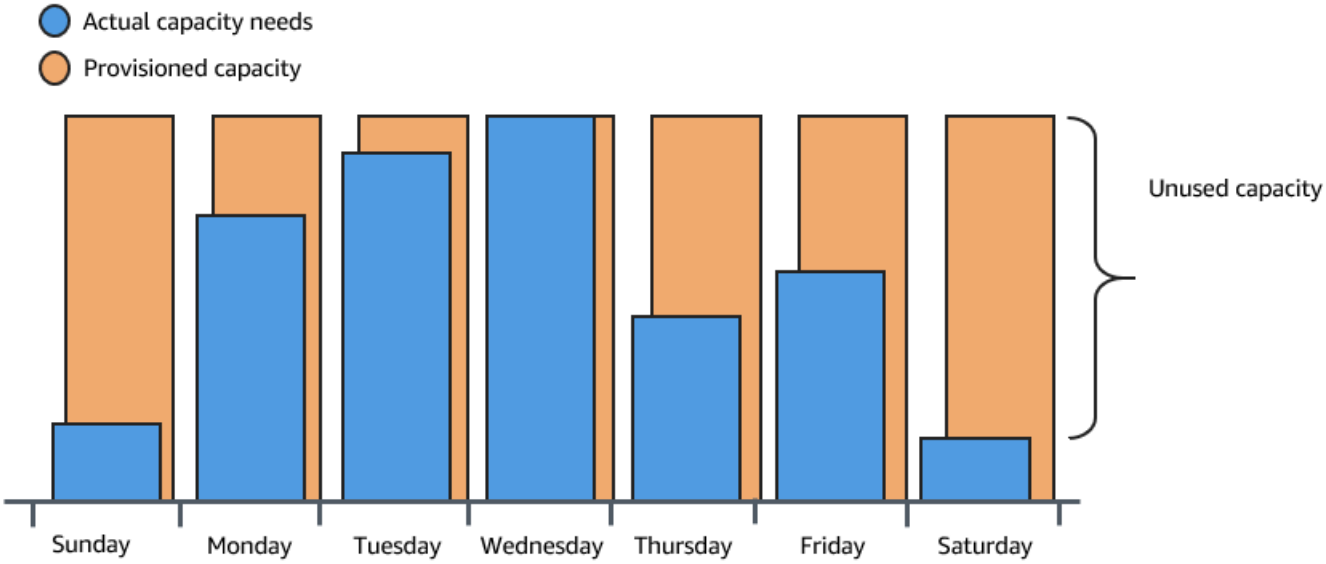
Exemplo: atender a demanda variável

Para demonstrar alguns dos benefícios do Amazon EC2 Auto Scaling, considere uma aplicação Web básica em execução na AWS. Essa aplicação permite que os funcionários pesquisem salas de conferência que podem usar para reuniões. Durante o início e o fim da semana, o uso dessa aplicação é mínimo. Durante o meio da semana, mais funcionários agendam reuniões, de forma que a demanda sobre a aplicação aumenta significativamente.

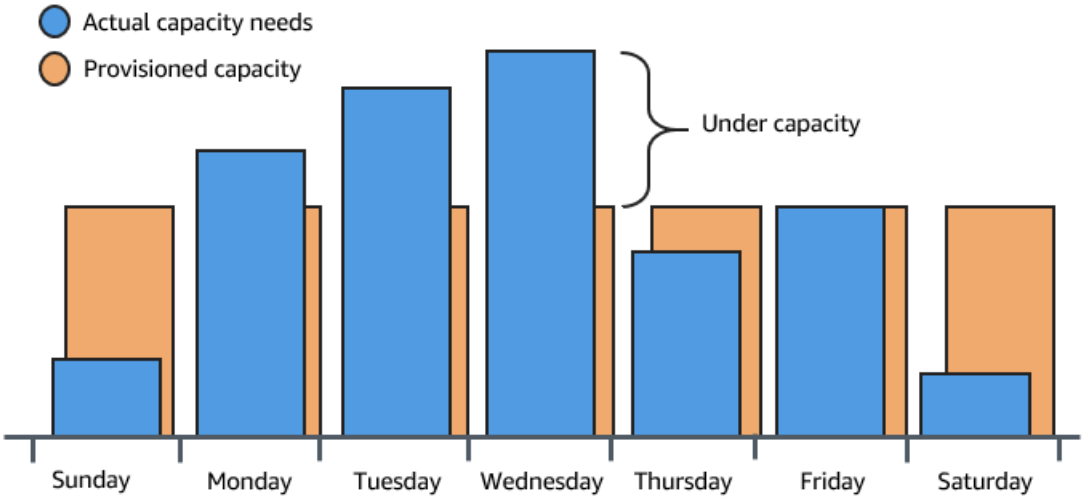
O gráfico a seguir mostra quanto da capacidade da aplicação é usado durante o período de uma semana.



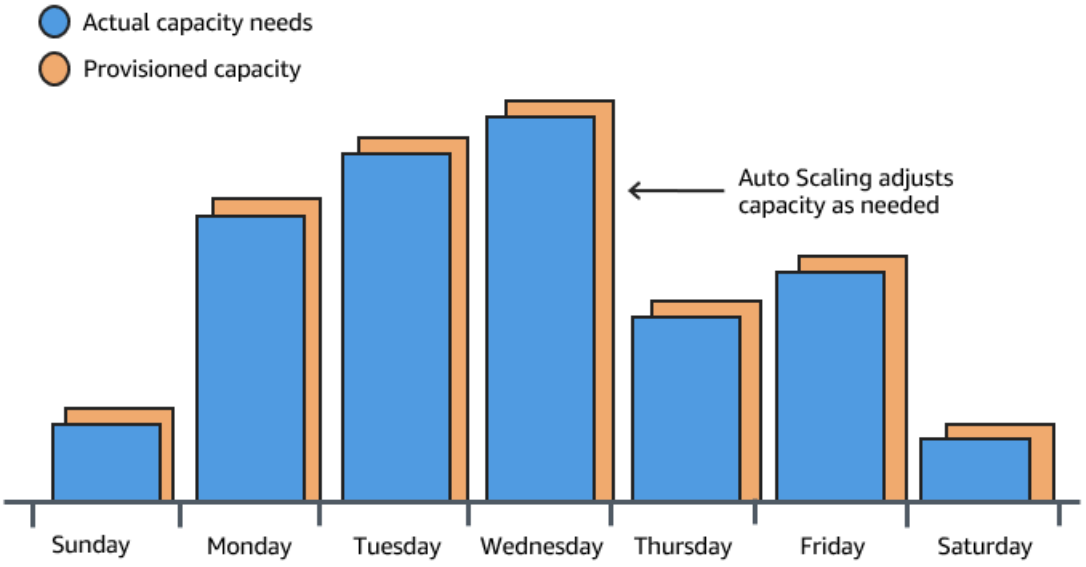
Tradicionalmente, há duas maneiras de planejar essas alterações na capacidade. A primeira opção é adicionar servidores suficientes para que a aplicação sempre tenha capacidade suficiente para atender à demanda. A desvantagem dessa opção, no entanto, é que há dias em que a aplicação não precisa de toda essa capacidade. A capacidade extra permanece não utilizada e, em essência, aumenta o custo de manutenção da aplicação em execução.



A segunda opção é ter capacidade suficiente para lidar com a demanda média na aplicação. Essa opção é mais barata, porque você não está comprando equipamento que usará apenas ocasionalmente. No entanto, você corre o risco de criar uma experiência do cliente insatisfatória quando a demanda na aplicação exceder sua capacidade.



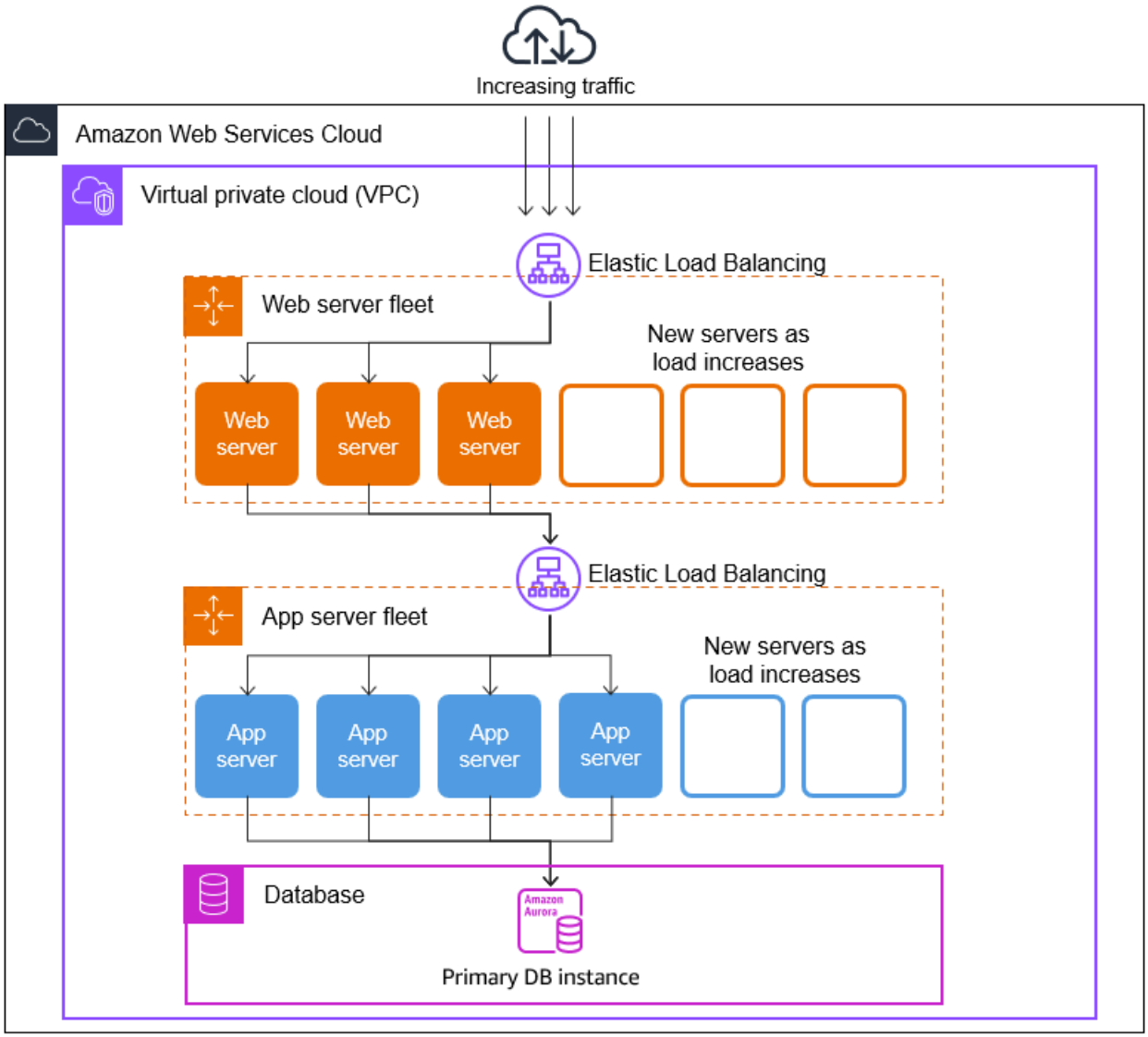
Ao adicionar o Amazon EC2 Auto Scaling a essa aplicação, você passa a ter uma terceira opção disponível. Você pode adicionar novas instâncias à aplicação somente quando necessário e encerrá-las quando não forem mais necessárias. Como o Amazon EC2 Auto Scaling usa instâncias do EC2, você só precisa pagar pelas instâncias que usa, quando as usa. Você agora tem uma arquitetura econômica que fornece a melhor experiência ao cliente e, ao mesmo tempo, minimiza os custos.



Exemplo: arquitetura de aplicação Web

Em um cenário comum de aplicação Web, você pode executar várias cópias da sua aplicação simultaneamente para cobrir o volume de tráfego de clientes. Essas várias cópias da aplicação são hospedadas em instâncias do EC2 idênticas (servidores de nuvem), cada uma lidando com solicitações de clientes.

O Amazon EC2 Auto Scaling gerencia a ativação e o encerramento dessas instâncias do EC2 em seu nome. Você define um conjunto de critérios (como um CloudWatch alarme da Amazon) que determina quando o grupo Auto Scaling inicia ou encerra instâncias do EC2. A adição de grupos do Auto Scaling à sua arquitetura de rede ajuda a tornar a aplicação mais disponível e tolerante a falhas.



Você pode criar tantos grupos do Auto Scaling quanto necessários. Por exemplo, você pode criar um grupo do Auto Scaling para cada camada.

Para distribuir o tráfego entre as instâncias em seus grupos do Auto Scaling, você pode inserir um balanceador de carga em sua arquitetura. Para obter mais informações, consulte Elastic Load Balancing.

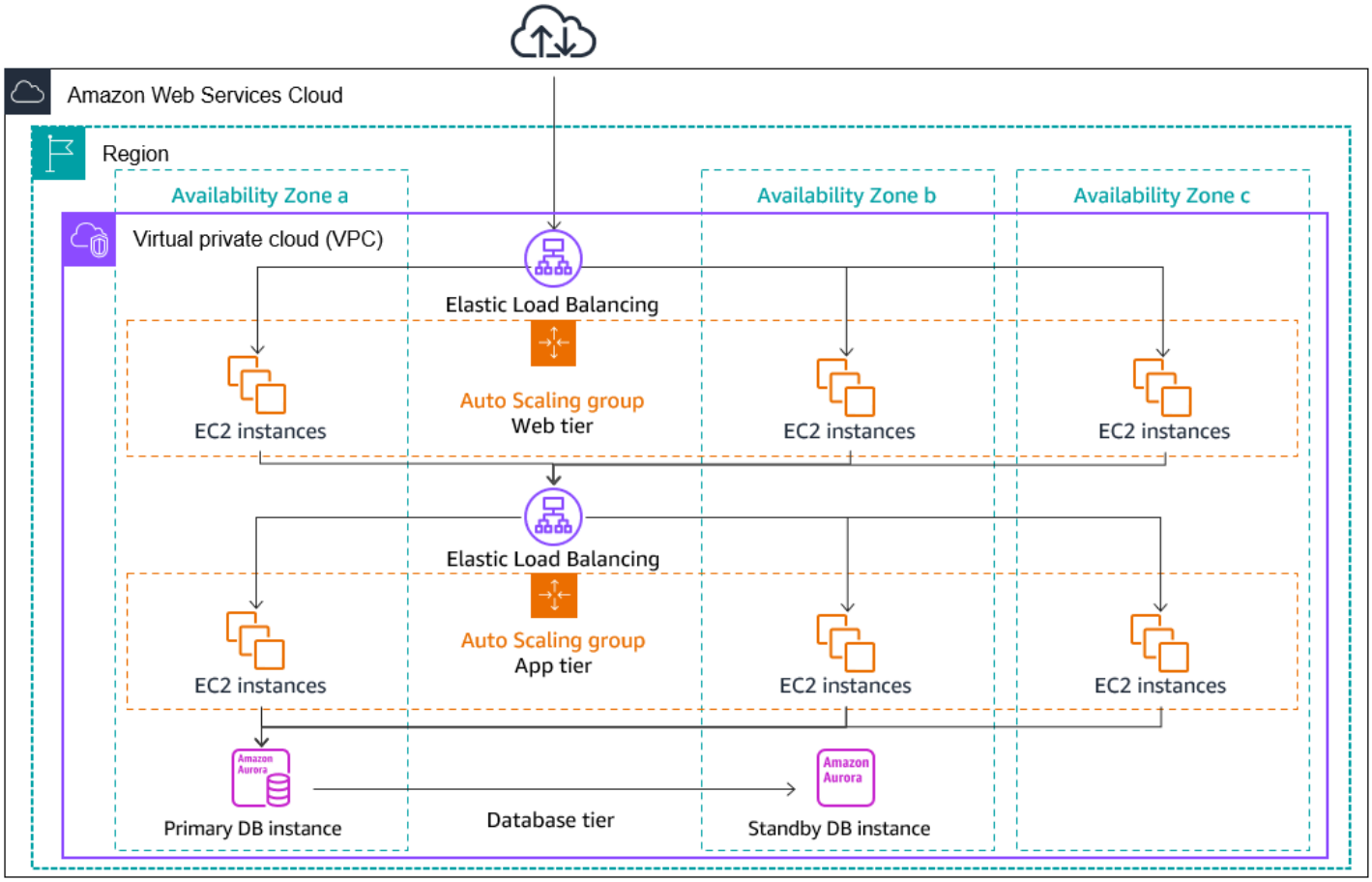
Exemplo: distribuir instâncias entre zonas de disponibilidade

As zonas de disponibilidade são locais isolados em uma determinada Região da AWS. Cada região tem várias zonas de disponibilidade, destinadas a fornecer alta disponibilidade para a região. As zonas de disponibilidade são independentes e, portanto, você aumenta a disponibilidade da aplicação quando a projeta para usar várias zonas. Para obter mais informações, consulte Resiliência no Amazon EC2 Auto Scaling.

Uma zona de disponibilidade é identificada pelo Região da AWS código seguido por um identificador de letra (por exemplo,`us-east-1a`). Se você criar a VPC e as sub-redes em vez de usar a VPC padrão, poderá definir uma ou mais sub-redes em cada zona de disponibilidade. Cada sub-rede deve residir inteiramente dentro de uma zona de disponibilidade e não pode abranger zonas. Para mais informações, consulte Como funciona a Amazon VPC no Manual do usuário da Amazon VPC.

Ao criar um grupo do Auto Scaling, é necessário escolher a VPC e as sub-redes nas quais implantará o grupo do Auto Scaling. O Amazon EC2 Auto Scaling cria as instâncias nas sub-redes escolhidas. Assim, cada instância é associada a uma zona de disponibilidade específica escolhida pelo Amazon EC2 Auto Scaling. Quando as instâncias são iniciadas, o Amazon EC2 Auto Scaling tenta distribuí-las uniformemente entre as zonas para garantir alta disponibilidade e confiabilidade.

A imagem a seguir mostra uma visão geral de uma arquitetura de vários níveis distribuída por três zonas de disponibilidade.



Distribuição de instâncias

O Amazon EC2 Auto Scaling tenta automaticamente manter números equivalentes de instâncias em cada zona de disponibilidade habilitada. O Amazon EC2 Auto Scaling faz isso tentando iniciar novas instâncias na zona de disponibilidade com o menor número de instâncias. Se houver várias sub-redes em uma zona de disponibilidade, o Amazon EC2 Auto Scaling tentará executar instâncias em uma sub-rede com o maior número de endereços IP disponíveis nessa zona de disponibilidade. No entanto, se a tentativa falhar, o Amazon EC2 Auto Scaling tentará iniciar as instâncias em outra zona de disponibilidade até obter êxito.

Em circunstâncias em que uma zona de disponibilidade perde a integridade ou deixa de estar disponível, a distribuição das instâncias entre as zonas de disponibilidade pode ficar desbalanceada. Quando a zona de disponibilidade se recupera, o Amazon EC2 Auto Scaling reequilibra automaticamente o grupo do Auto Scaling. Ele faz isso iniciando instâncias nas zonas de disponibilidade habilitadas que têm menos instâncias e encerrando as instâncias em outros locais.

Atividades de rebalanceamento

As atividades de rebalanceamento dividem-se em duas categorias: rebalanceamento de zona de disponibilidade e rebalanceamento de capacidade.

Rebalanceamento de zona de disponibilidade

Após determinadas ações ocorrerem, seu grupo do Auto Scaling poderá se tornar desbalanceado entre as zonas de disponibilidade. O Amazon EC2 Auto Scaling compensará rebalanceando as zonas de disponibilidade. As ações a seguir podem levar a atividade de rebalanceamento:

- Você altera as zonas de disponibilidade associadas ao grupo do Auto Scaling.
- Você explicitamente encerra ou desanexa instâncias, ou as coloca em espera e assim o grupo fica desbalanceado.
- Uma zona de disponibilidade que antes tinha capacidade insuficiente se recupera e passa a ter capacidade adicional.
- Uma zona de disponibilidade que tinha um preço spot acima do seu preço spot máximo agora tem um preço spot abaixo do seu preço máximo.

Ao rebalancear instâncias, o Amazon EC2 Auto Scaling inicia novas instâncias antes de encerrar as mais antigas. Dessa forma, o rebalanceamento não compromete a performance nem a disponibilidade da aplicação.

Como o Amazon EC2 Auto Scaling tenta iniciar novas instâncias antes de encerrar as mais antigas, estar usando toda ou quase toda a capacidade máxima especificada pode prejudicar ou parar completamente as atividades de rebalanceamento.

Para evitar esse problema, o sistema pode exceder temporariamente a capacidade máxima especificada de um grupo durante uma atividade de rebalanceamento. Por padrão, isso pode ser feito com uma margem de 10% ou uma instância, o que for maior. A margem só é estendida se o grupo estiver usando toda ou quase toda a capacidade máxima e precisar ser rebalanceado. A extensão dura somente o tempo necessário para rebalancear o grupo (em geral, alguns minutos).

Como alternativa, você pode estabelecer limites para um grupo do Auto Scaling usando uma política de manutenção de instâncias, e o grupo só pode aumentar ou diminuir a capacidade dentro dessa faixa de limite. Dessa forma, você pode controlar a rapidez com que seu grupo se reequilibra. Para obter mais informações, consulte Políticas de manutenção de instância.

Rebalanceamento de capacidade

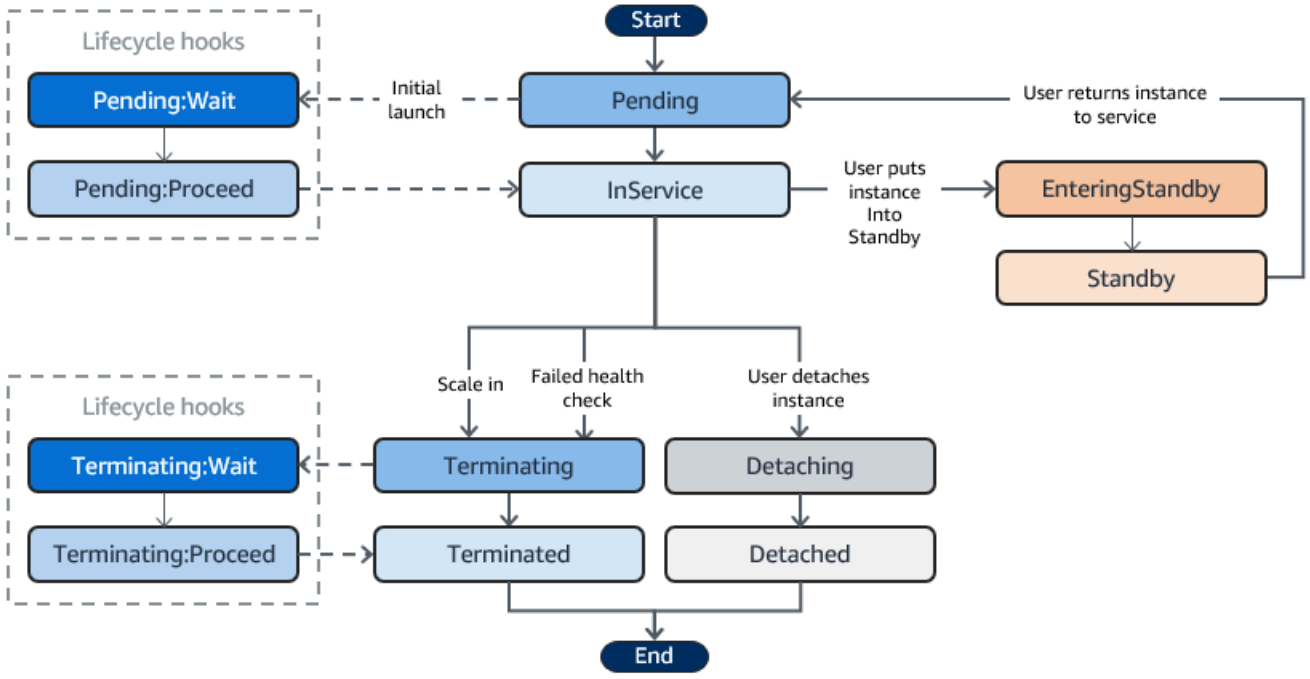
Você pode habilitar o rebalanceamento de capacidade nos grupos do Auto Scaling usando instâncias spot. O Amazon EC2 Auto Scaling tenta iniciar uma instância spot sempre que o Amazon EC2 informa que uma instância spot está em alto risco de ser interrompida. Após iniciar uma nova instância, ele encerra uma instância mais antiga. Para obter mais informações, consulte Rebalanceamento de capacidade no Auto Scaling para substituir instâncias spot em risco.

Ciclo de vida das instâncias do Amazon EC2 Auto Scaling

As instâncias do EC2 em um grupo do Auto Scaling têm um caminho ou um ciclo de vida que difere daquele de outras instâncias do EC2. O ciclo de vida começa quando o grupo do Auto Scaling ativa uma instância e a coloca em serviço. O ciclo de vida termina quando você encerra a instância, ou o grupo do Auto Scaling retira a instância de serviço e a termina.

> **Note**
>
> Você é cobrado pelas instâncias assim que elas são ativadas, incluindo o tempo em que elas ainda não estão em serviço.

A ilustração a seguir mostra as transições entre estados de instâncias no ciclo de vida do Amazon EC2 Auto Scaling.



Escalonamento horizontal

Os seguintes eventos de aumento da escala na horizontal instruem o grupo do Auto Scaling a iniciar instâncias do EC2 e anexá-las ao grupo:

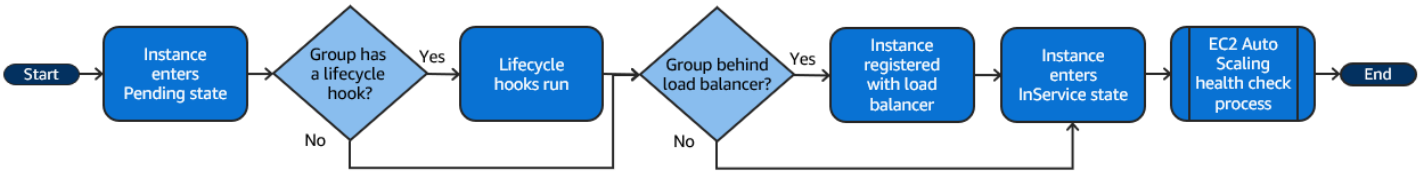
- Você aumenta o tamanho do grupo manualmente. Para obter mais informações, consulte Alterar a capacidade desejada de um grupo do Auto Scaling existente.
- Você cria uma política de escalabilidade para aumentar automaticamente o tamanho do grupo com base em um aumento especificado na demanda. Para obter mais informações, consulte Escalabilidade dinâmica para o Amazon EC2 Auto Scaling.
- Você configura a escalabilidade programando o aumento do tamanho do grupo em um horário específico. Para obter mais informações, consulte Escalabilidade programada para o Amazon EC2 Auto Scaling.

Quando um evento para aumentar a escala na horizontal ocorre, o grupo do Auto Scaling executa o número necessário de instâncias do EC2 usando seu modelo de execução atribuído. Essas instâncias iniciam no estado `Pending`. Se adicionar um gancho do ciclo de vida a seu grupo do Auto Scaling, você poderá executar uma ação personalizada aqui. Para obter mais informações, consulte Hooks do ciclo de vida.

Quando cada instância está totalmente configurada e passa nas verificações de integridade do Amazon EC2, elas são anexadas ao grupo do Auto Scaling e entram no estado `InService`. A instância é contabilizada para a capacidade desejada do grupo do Auto Scaling.

Se o grupo do Auto Scaling estiver configurado para receber tráfego de um balanceador de carga do Elastic Load Balancing, o Amazon EC2 Auto Scaling registrará automaticamente a instância no balanceador de carga antes de marcar a instância como `InService`.

A seguir, resumem-se as etapas para registrar uma instância com um balanceador de carga para um evento de aumento de escala.



Instâncias em serviço

As instâncias permanecem no estado `InService` até que ocorra um dos seguintes eventos:

- Um evento de redução da escala na horizontal ocorre e o Amazon EC2 Auto Scaling escolhe terminar essa instância para reduzir o tamanho do grupo do Auto Scaling. Para obter mais informações, consulte Controlar quais instâncias do Auto Scaling serão terminadas durante uma redução de escala na horizontal.
- Você coloca a instância em um estado `Standby`. Para obter mais informações, consulte Entrar e sair de espera.
- Você desvincula a instância do grupo do Auto Scaling. Para obter mais informações, consulte Desvincular instâncias do seu grupo do Auto Scaling.
- A instância não é aprovada em um número necessário de verificações de integridade e, portanto, é removida do grupo do Auto Scaling, terminada e substituída. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Reduzir a escala na horizontal

Os seguintes eventos de redução da escala na horizontal instruem o grupo do Auto Scaling a desvincular instâncias do EC2 do grupo e a encerrá-las:

- Você reduz o tamanho do grupo manualmente. Para obter mais informações, consulte Alterar a capacidade desejada de um grupo do Auto Scaling existente.
- Você cria uma política de escalabilidade para reduzir automaticamente o tamanho do grupo com base em uma redução especificada na demanda. Para obter mais informações, consulte Escalabilidade dinâmica para o Amazon EC2 Auto Scaling.
- Você configura a escalabilidade programando a redução do tamanho do grupo em um horário específico. Para obter mais informações, consulte Escalabilidade programada para o Amazon EC2 Auto Scaling.

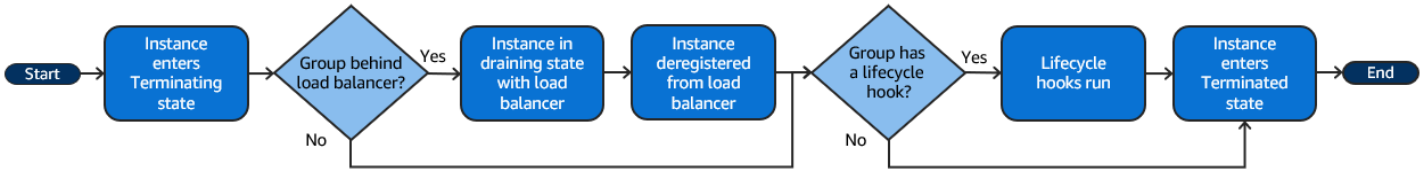
É importante criar um evento de redução correspondente para cada evento de expansão criado. Isso ajuda a garantir que os recursos atribuídos à aplicação correspondam à demanda por esses recursos da maneira mais próxima possível.

Quando um evento de redução da escala na horizontal ocorre, o grupo do Auto Scaling termina uma ou mais instâncias. O grupo do Auto Scaling usa sua política de término para determinar quais instâncias devem ser terminadas. As instâncias que estão em processo de encerramento do grupo do Auto Scaling entram no `Terminating` estado e não podem ser colocadas novamente em serviço.

Se o seu grupo do Auto Scaling estiver configurado para receber tráfego de um load balancer do Elastic Load Balancing, o Amazon EC2 Auto Scaling cancelará automaticamente o registro da instância final do load balancer. O cancelamento do registro da instância garante que todas as novas solicitações sejam redirecionadas para outras instâncias no grupo de destino do balanceador de carga, enquanto as conexões existentes com a instância podem continuar até que o atraso de cancelamento de registro expire.

Se você adicionar um gancho do ciclo de vida ao grupo do Auto Scaling, poderá executar uma ação personalizada na instância final. Para obter mais informações, consulte Hooks do ciclo de vida. Finalmente, a instância é completamente encerrada e entra no estado`Terminated`.

A seguir, um resumo dos passos para desregistrar uma instância de um balanceador de carga em um evento de diminuição de escala.



Desvincular uma instância

Você pode desvincular uma instância do seu grupo do Auto Scaling. Depois que a instância for desvinculada, você poderá gerenciá-la separadamente do grupo do Auto Scaling ou anexá-la a outro grupo do Auto Scaling.

Para obter mais informações, consulte Desvincular instâncias do seu grupo do Auto Scaling.

Anexar uma instância

Você pode anexar uma instância do EC2 em execução que atenda a determinados critérios a seu grupo do Auto Scaling. Após ser anexada, a instância é gerenciada como parte do grupo do Auto Scaling.

Para obter mais informações, consulte Desvincular instâncias do seu grupo do Auto Scaling.

Hooks do ciclo de vida

Você pode adicionar um gancho do ciclo de vida ao grupo do Auto Scaling para ativar ações personalizadas quando as instâncias forem iniciadas ou terminadas.

Quando o Amazon EC2 Auto Scaling responde a um evento de aumento da escala na horizontal, ele inicia uma ou mais instâncias. Essas instâncias iniciam no estado `Pending`. Se você adicionar um gancho do ciclo de vida `autoscaling:EC2_INSTANCE_LAUNCHING` ao grupo do Auto Scaling, as instâncias avançarão do estado `Pending` para o estado `Pending:Wait`. Depois que você concluir a ação do ciclo de vida, as instâncias entrarão no estado `Pending:Proceed`. Quando as instâncias estão totalmente configuradas, elas são anexadas ao grupo do Auto Scaling e entram no estado `InService`.

Quando o Amazon EC2 Auto Scaling responde a um evento de redução da escala na horizontal, ele encerra uma ou mais instâncias. Essas instâncias são desvinculadas do grupo do Auto Scaling e entram no estado `Terminating`. Se você adicionar um gancho do ciclo de vida `autoscaling:EC2_INSTANCE_TERMINATING` ao grupo do Auto Scaling, as instâncias avançarão do estado `Terminating` para o estado `Terminating:Wait`. Depois que você concluir a ação do ciclo de vida, as instâncias entrarão no estado `Terminating:Proceed`. Quando as instâncias estão totalmente encerradas, elas entram no estado `Terminated`.

Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

Entrar e sair de espera

Você pode colocar qualquer instância que esteja em um estado `InService` em um estado `Standby`. Isso permite que você remova a instância de serviço, solucione problemas ou faça alterações na instância e coloque-a em serviço novamente.

As instâncias em estado `Standby` continuam a ser gerenciadas pelo grupo do Auto Scaling. No entanto, elas não fazem parte ativamente da aplicação até que você as coloque em serviço novamente.

Para obter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.

Cotas para recursos e grupos do Auto Scaling

Você Conta da AWS tem cotas padrão, anteriormente chamadas de limites, para cada AWS serviço. A menos que especificado de outra forma, cada cota é específica da região . Você pode solicitar o aumento de algumas cotas, porém, algumas delas não podem ser aumentadas.

Para visualizar as cotas do Amazon EC2 Auto Scaling, abra o console do Service Quotas. No painel de navegação, escolha AWS services (serviços) e selecione Amazon EC2 Auto Scaling.

Para solicitar o aumento da cota, consulte Solicitar um aumento de cota no Guia do usuário do Service Quotas. Se a cota ainda não estiver disponível em Service Quotas, use o Auto Scaling limits form (Formulário de limites do Auto Scaling). Os aumentos de cota estão vinculados à região para a qual são solicitados.

Recursos do Amazon EC2 Auto Scaling

Você Conta da AWS tem as seguintes cotas relacionadas ao número de grupos de Auto Scaling e configurações de lançamento que você pode criar.

Recurso	Cota padrão
Grupos do Auto Scaling por região	500
Configuração de execução por região	200

Configuração do grupo do Auto Scaling

Você Conta da AWS tem as seguintes cotas relacionadas à configuração dos grupos do Auto Scaling. Eles não podem ser alterados.

Recurso	Quota
Políticas de escalabilidade por grupo do Auto Scaling	50
Ações programadas por grupo do Auto Scaling	125
Ajustes de etapa por política de escalabilidade de etapa	20
Hooks do ciclo de vida por grupo do Auto Scaling	50
Tópicos do SNS por grupo do Auto Scaling	10
Classic Load Balancers por grupo do Auto Scaling	50
Grupos de destino do Elastic Load Balancing por grupo do Auto Scaling	50
Grupos de destino do VPC Lattice por grupo do Auto Scaling	5

Operações da API do grupo do Auto Scaling

O Amazon EC2 Auto Scaling fornece operações de API para fazer alterações em seus grupos do Auto Scaling em lotes. Veja a seguir os limites da API no número máximo de itens (máximo de membros da matriz) permitidos em uma única operação. Eles não podem ser alterados.

Operation	Máximo de membros da matriz
AttachInstances	20 instâncias IDs
AttachLoadBalancers	10 balanceadores de cargas
AttachLoadBalancerTargetGroups	10 grupos de destino
BatchDeleteScheduledAction	50 ações programadas
BatchPutScheduledUpdateGroupAction	50 ações programadas
DetachInstances	20 instâncias IDs
DetachLoadBalancers	10 balanceadores de cargas
DetachLoadBalancerTargetGroups	10 grupos de destino
EnterStandby	20 instâncias IDs
ExitStandby	20 instâncias IDs
SetInstanceProtection	50 instâncias IDs

Limitação de controle de utilização para a API do Amazon EC2 Auto Scaling

A utilização da API do Amazon EC2 Auto Scaling e controlada usando um esquema de bucket do token para manter a largura de banda do serviço. Para mais informações, consulte Taxa de solicitação da API na Referência da API do Amazon EC2 Auto Scaling.

Taxas de encerramento do EC2

O Amazon EC2 Auto Scaling determina dinamicamente o número de operações de encerramento da instância do EC2 que podem ser executadas por vez quando seu grupo do Auto Scaling sofrer redução. Isso significa que você pode ver variações no número de instâncias encerradas por vez nos grupos do Auto Scaling. Essas variações são causadas por considerações externas, como se o Amazon EC2 Auto Scaling precisasse cancelar o registro de instâncias com um balanceador de carga.

Outros produtos da

As cotas para outros serviços, como o Amazon EC2 e o Amazon VPC, podem afetar seus grupos do Auto Scaling. Você pode usar Service Quotas para atualizar as cotas para instâncias do EC2 e outros recursos em seu. Conta da AWS No Service Quotas console, você pode ver todas as cotas de serviço disponíveis e solicitar aumentos para elas. Para obter mais informações, consulte Solicitar um aumento de cota no Guia do usuário Service Quotas .

Para obter cotas específicas para modelos de execução, consulte Restrições de modelos de execução no Guia do usuário do Amazon EC2.

Configurar para usar o Amazon EC2 Auto Scaling

Antes de começar a usar o Amazon EC2 Auto Scaling, conclua as tarefas a seguir.

Tarefas

- Prepare-se para usar o AWS CLI

Prepare-se para usar o AWS CLI

Você pode usar as ferramentas de linha de AWS comando para emitir comandos na linha de comando do seu sistema para realizar o Amazon EC2 Auto Scaling e outras tarefas. AWS

Para usar o AWS Command Line Interface (AWS CLI), baixe, instale e configure a versão 1 ou 2 do AWS CLI. A mesma funcionalidade do Amazon EC2 Auto Scaling está disponível nas versões 1 e 2. Para instalar a versão 1 AWS CLI , consulte Instalar, atualizar e desinstalar a AWS CLI no Guia do usuário da AWS CLI versão 1. Para instalar a AWS CLI versão 2, consulte Instalando ou atualizando a versão mais recente do AWS CLI no Guia do Usuário da AWS CLI Versão 2.

AWS CloudShell permite que você pule a instalação do AWS CLI em seu ambiente de desenvolvimento e, em Console de gerenciamento da AWS vez disso, o use no. Além de evitar a instalação, não é necessário configurar credenciais nem especificar uma região. Sua Console de gerenciamento da AWS sessão fornece esse contexto para AWS CLI o. Você pode usar AWS CloudShell em suportado Regiões da AWS. Para obter mais informações, consulte Crie grupos de Auto Scaling a partir da linha de comando usando AWS CloudShell.

Para obter mais informações, consulte escalabilidade automática na Referência de comandos da AWS CLI .

Conceitos básicos do Amazon EC2 Auto Scaling

Para começar a usar o Amazon EC2 Auto Scaling, você pode seguir os tutoriais que apresentam o serviço.

Tópicos

- Tutorial: Criar seu primeiro grupo do Auto Scaling
- Tutorial: Configurar uma aplicação escalonada e com balanceamento de carga

Para obter tutoriais adicionais que se concentram em ferramentas específicas para gerenciar o ciclo de vida das instâncias em um grupo do Auto Scaling, consulte os seguintes tópicos:

- Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda. Este tutorial mostra como usar a Amazon EventBridge para criar regras que invocam funções Lambda com base em eventos que acontecem com as instâncias em seu grupo de Auto Scaling.
- Tutorial: Usar script de dados e metadados de instância para recuperar o estado do ciclo de vida. Este tutorial mostra como usar o Serviço de metadados de instância (IMDS) para invocar uma ação de dentro da própria instância.

Antes de criar um grupo do Auto Scaling para usar com sua aplicação, analise detalhadamente sua aplicação ao executá-la na Nuvem AWS. Considere o seguinte:

- Quantas zonas de disponibilidade o grupo do Auto Scaling deve abranger.
- Quais recursos existentes podem ser usados, como grupos de segurança ou Amazon Machine Images (AMIs).
- Se você deseja dimensionar para aumentar ou diminuir a capacidade ou se deseja apenas garantir que um número específico de servidores esteja sempre em execução. Lembre-se de que o Amazon EC2 Auto Scaling pode fazer as duas coisas simultaneamente.
- Quais métricas têm mais relevância para a performance da aplicação.
- Quanto tempo é necessário para iniciar e provisionar um servidor.

Quanto melhor você entender sua aplicação, mais eficaz você pode tornar sua arquitetura de Auto Scaling.

Tutorial: Criar seu primeiro grupo do Auto Scaling

Este tutorial fornece uma introdução prática ao Amazon EC2 Auto Scaling por meio do. Console de gerenciamento da AWS Você criará um modelo de inicialização que define suas instâncias do EC2 e um grupo do Auto Scaling com uma única instância. Após lançar seu grupo de Auto Scaling, você encerrará a instância e verificará se a instância foi removida do serviço e substituída. Para manter um número constante de instâncias, o Amazon EC2 Auto Scaling detecta e responde automaticamente às verificações de integridade e acessibilidade do Amazon EC2.

Ao se inscrever AWS, você pode começar a usar o Amazon EC2 Auto Scaling gratuitamente AWS usando o nível gratuito. É possível usar o nível gratuito para iniciar e usar uma instância `t2.micro` gratuitamente por 12 meses (em regiões onde `t2.micro` não estiver disponível, será possível usar uma instância `t3.micro` no nível gratuito). Se você iniciar uma instância que não esteja no nível gratuito, serão cobradas as taxas de uso padrão do Amazon EC2 para a instância. Para obter mais informações, consulte Definição de preço do Amazon EC2.

Tarefas

- Preparar para a demonstração
- Etapa 1: Criar um modelo de execução
- Etapa 2: Criar um grupo do Auto Scaling com uma única instância
- Etapa 3: Verificar seu grupo do Auto Scaling
- Etapa 4: Terminar uma instância no seu grupo do Auto Scaling
- Etapa 5: Próximas etapas
- Etapa 6: limpar

Preparar para a demonstração

Este passo a passo pressupõe que você esteja familiarizado com a execução de instâncias do EC2 e que já criou um par de chaves e um grupo de segurança.

SPara começar a usar o Amazon EC2 Auto Scaling, você pode usar a VPC padrão para o seu Conta da AWS. A VPC padrão inclui uma sub-rede pública padrão em cada zona de disponibilidade e um gateway de Internet conectado à VPC. Você pode visualizá-lo VPCs na VPCspágina Seu do console da Amazon Virtual Private Cloud (Amazon VPC).

Etapa 1: Criar um modelo de execução

Nesta etapa, você cria um modelo de execução que especifica o tipo de instância EC2 do Amazon EC2 Auto Scaling que ele cria para você. Inclua informações, como o ID da imagem de máquina da Amazon (AMI) a ser usada, o tipo de instância, o par de chaves e os grupos de segurança.

Para criar um modelo de execução

1. Abra Launch templates (Modelos de execução) no console do Amazon EC2.
2. Na barra de navegação superior, selecione um Região da AWS. O modelo de execução e os recursos do grupo do Auto Scaling que você cria são vinculados à região que você especifica.
3. Escolha Criar modelo de execução.
4. Para o Launch template name (Nome do modelo de execução), insira **`my-template-for-auto-scaling`**.
5. Em Auto Scaling guidance (Guia do Auto Scaling), marque a caixa de seleção.
6. Em Application and OS Images (Amazon Machine Image) (Imagens de aplicações e sistemas operacionais [imagem de máquina da Amazon]), escolha uma versão do Amazon Linux 2 (HVM) na lista Quick Start (Início rápido). A AMI serve como modelo de configuração básico para suas instâncias.
7. Em Instance type (Tipo de instância), selecione uma configuração de hardware que seja compatível com a AMI que você especificou.
8. (Opcional) Em Key pair (login) (Par de chaves [login]), escolha um par de chaves existente. Você usa pares de chaves para se conectar a uma instância do Amazon EC2 co o SSH. A conexão a uma instância não está incluída como parte deste tutorial. Portanto, não é necessário especificar um par de chaves, a menos que pretenda se conectar à instância usando SSH.
9. Em Network settings (Configurações de rede), expanda Advanced network configuration (Configuração de rede avançada) e execute estas ações:

    a. Escolha Add network interface (Adicionar interface de rede) para configurar a interface de rede primária.

    b. Para o Auto-assign Public IP (Autoatribuir IP público): especifique se sua instância recebe um endereço IPv4 público. Por padrão, o Amazon EC2 atribui um endereço IPv4 público se a instância do EC2 for iniciada em uma sub-rede padrão ou se a instância for iniciada em uma sub-rede configurada para atribuir automaticamente um endereço público. IPv4 Não escolha se você precisar conectar-se à sua instância. escolha Desabilitar

c. Para IDs deSecurity groups (Grupos de segurança), especifique o grupo de segurança na mesma VPC que planeja usar como a VPC para o grupo do Auto Scaling. Se você não especificar um grupo de segurança, sua instância será automaticamente associada ao grupo de segurança padrão da VPC.

d. Para Delete on termination (Excluir no encerramento):, escolha Sim para excluir a interface de rede quando a instância for deletada.

10. Escolha Criar modelo de execução.
11. Na página de confirmação, escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Etapa 2: Criar um grupo do Auto Scaling com uma única instância

Use o seguinte procedimento para continuar de onde você parou após criar um modelo de lançamento.

Para criar um grupo do Auto Scaling

1. Na página Choose launch template or configuration (Escolher modelo ou configuração de execução), em Auto Scaling group name (Nome do grupo do Auto Scaling), insira **`my-first-asg`**.
2. Escolha Próximo.

   A página Escolher opções de lançamento da instância aparece, permitindo que você escolha as configurações de rede da VPC que deseja que o grupo de Auto Scaling utilize, além de oferecer opções para lançar instâncias sob demanda e spot.
3. Na seção Rede, mantenha a VPC definida como a VPC padrão de sua escolha ou selecione sua própria Região da AWS VPC. A VPC padrão é configurada automaticamente para fornecer conectividade com a Internet à sua instância. Essa VPC inclui uma sub-rede pública em cada zona de disponibilidade na região.
4. Para Availability Zones and subnets (Zonas de disponibilidade e sub-redes), escolha uma sub-rede de cada zona de disponibilidade que você desejar incluir. Use sub-redes em várias zonas de disponibilidade para alta disponibilidade. Para obter mais informações, consulte Considerações sobre a escolha de sub-redes da VPC.
5. Na seção Instance type requirements (Requisitos de tipo de instância), use a configuração padrão para simplificar essa etapa. (Não substitua o modelo de execução.) Neste tutorial, você fará o execução de apenas uma das Instâncias sob demanda usando o tipo de instância especificado no modelo de execução.

6. Mantenha o restante dos padrões para este tutorial e escolha Skip to review (Avançar para a revisão).

> **Note**
>
> O tamanho inicial do grupo é determinado pela capacidade desejada. O valor padrão é uma instância 1.

7. Em Review (Revisar), analise as informações do grupo e selecione Create Auto Scaling group (Criar grupo do Auto Scaling).

Etapa 3: Verificar seu grupo do Auto Scaling

Agora que criou seu grupo do Auto Scaling, você está pronto para verificar se o grupo iniciou uma instância do EC2.

> **Tip**
>
> No procedimento a seguir, você visualiza as seções Activity history (Histórico de atividades) e Instances (Instâncias) do grupo do Auto Scaling. Em ambas, as colunas nomeadas já deverão ser exibidas. Para exibir colunas ocultas ou alterar o número de linhas exibidas, escolha o ícone de engrenagem, no canto superior direito de cada seção, para abrir o modal de preferências, atualize as configurações conforme necessário e escolha Confirm (Confirmar).

Para verificar se seu grupo do Auto Scaling iniciou uma instância do EC2

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling recém-criado.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling). A primeira guia disponível é a guia Details (Detalhes), que mostra informações sobre o grupo do Auto Scaling.
3. Escolha a segunda guia, Activity (Atividade). Em Activity history (Histórico de atividades), é possível visualizar o progresso das atividades associadas ao grupo do Auto Scaling. A coluna Status mostra o status atual de sua instância. Enquanto sua instância está ativando, a coluna de

status mostra `Not yet in service`. O status muda para `Successful` depois que a instância é ativada. Você também pode usar o botão Atualizar para ver o status atual de sua instância.

4. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), é possível visualizar o status da instância.
5. Verifique se sua instância foi executada com êxito. Demora um pouco para iniciar uma instância.

   - A guia Lifecycle (Ciclo de vida) mostra o estado de sua instância. Inicialmente, sua instância está no estado `Pending`. Quando uma instância está pronta para receber tráfego, seu estado é `InService`.
   - A coluna Status de saúde mostra o resultado das verificações de saúde do Amazon EC2 Auto Scaling na sua instância.

Etapa 4: Terminar uma instância no seu grupo do Auto Scaling

Use estas etapas para saber mais sobre como o Amazon EC2 Auto Scaling funciona, especificamente, como ele executa novas instâncias quando necessário. O tamanho mínimo para o grupo do Auto Scaling criado neste tutorial é de uma instância. Portanto, se você terminar essa instância em execução, o Amazon EC2 Auto Scaling deverá iniciar uma nova instância para substituí-la.

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.
3. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), selecione o ID da instância.

   Isso o levará até a página Instances (Instâncias) do console do Amazon EC2, onde é possível encerrar a instância.
4. Escolha Actions (Ações), Instance State (Estado da instância), Terminate (Encerrar). Quando a confirmação for solicitada, escolha Sim, encerrar.
5. No painel de navegação, em Auto Scaling, escolha Auto Scaling Groups (Grupos de Auto Scaling). Selecione seu grupo do Auto Scaling e escolha a guia Activity (Atividade).

   Quando você encerra uma instância na página Instâncias, leva um ou dois minutos após o encerramento da instância para que uma nova instância seja iniciada. No histórico de atividades, quando a ação de escalabilidade for iniciada, você observará uma entrada para o encerramento

da primeira instância e uma entrada para a execução de uma nova instância. Use o botão de atualização até ver as novas entradas.

6. Na guia Instance management (Gerenciamento de instâncias), a seção Instances (Instâncias) exibe somente a nova instância.
7. No painel de navegação, em Instances (Instâncias), escolha Instances (Instâncias). Essa página mostra a instância encerrada e a instância em execução.

Etapa 5: Próximas etapas

Vá para a próxima etapa se quiser excluir a infraestrutura básica que você acabou de criar. Caso contrário, você pode usar essa infraestrutura como sua base e experimentar uma ou mais das seguintes:

- Conectar-se à sua instância do Linux usando o Gerenciador de sessões ou o SSH. Para obter mais informações, consulte Conexão à sua instância do EC2 usando o Session Manager e Conexão à sua instância do Linux usando o SSH no Guia do usuário do Amazon EC2.
- Configure uma notificação do Amazon SNS para notificar você sempre que seu grupo do Auto Scaling iniciar ou terminar instâncias. Para obter mais informações, consulte Notificações do Amazon SNS.
- Escalar manualmente seu grupo do Auto Scaling para testar a notificação do SNS. Para obter mais informações, consulte Alterar a capacidade desejada de seu grupo do Auto Scaling.

Você também pode começar a se familiarizar com os conceitos de escalonamento lendo sobre Políticas de escalabilidade de rastreamento de destino. Se a carga da sua aplicação mudar, seu grupo do Auto Scaling poderá aumentar a escala horizontalmente (adicionar instâncias) ou reduzir a escala horizontalmente (executar menos instâncias) automaticamente ajustando a capacidade desejada do grupo entre os limites mínimo e máximo de capacidade. Para obter mais informações sobre esses limites, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.

Etapa 6: limpar

Você pode excluir sua infraestrutura de escalabilidade ou excluir apenas o grupo do Auto Scaling e manter o modelo de execução para usar em outro momento.

Se você executou uma instância que não está no nível gratuito da AWS, é necessário terminar sua instância para evitar cobranças adicionais. Ao encerrar a instância, os dados associados a ela também serão excluídos.

Para excluir seu grupo do Auto Scaling

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling (`my-first-asg`).
3. Escolha Excluir.
4. Quando a confirmação for solicitada, digite **`delete`** para confirmar a exclusão do grupo do Auto Scaling especificado e, em seguida, escolha Excluir.

   Um ícone de carregamento na coluna Name (Nome) indica que o grupo do Auto Scaling está sendo excluído. Quando a exclusão tiver ocorrido, as colunas Desejado, Mínimo e Máximo exibirão `0` instâncias para o grupo do Auto Scaling. São necessários alguns minutos para encerrar a instância e excluir o grupo. Atualize a lista para ver o estado atual.

Ignore esse procedimento se quiser manter seu modelo de execução.

Para excluir seu modelo de execução

1. Abra a página Launch templates (Modelos de execução) do console do Amazon EC2.
2. Selecione o modelo de execução (`my-template-for-auto-scaling`).
3. Escolha Actions (Ações), Delete template (Excluir modelo).
4. Quando a confirmação for solicitada, digite **`Delete`** para confirmar a exclusão do modelo de execução especificado e, em seguida, escolha Excluir.

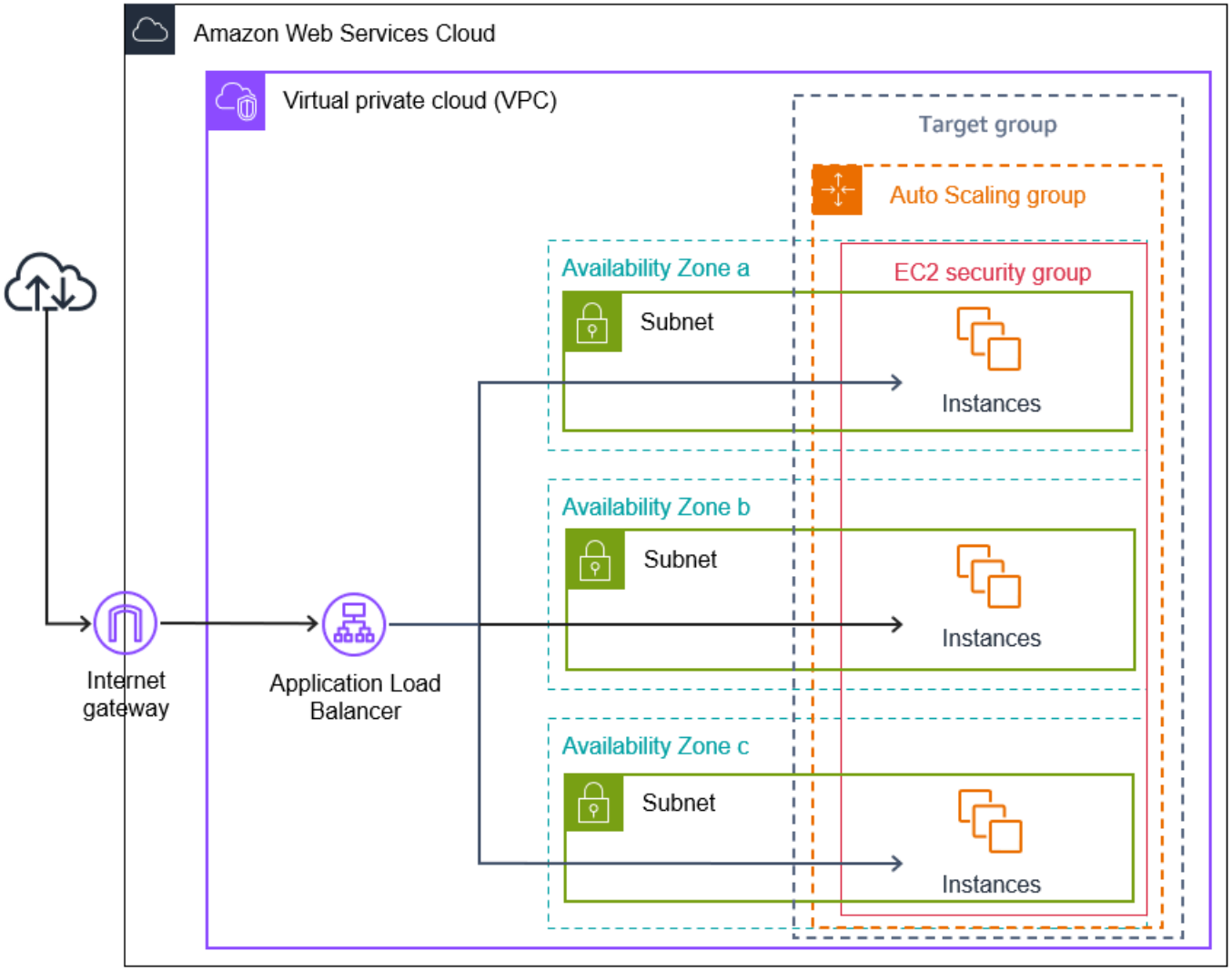
Tutorial: Configurar uma aplicação escalonada e com balanceamento de carga

> **Important**
>
> Antes de explorar este tutorial, recomendamos que você primeiramente examine o tutorial introdutório: Crie seu primeiro grupo do Auto Scaling.

O registro do seu grupo do Auto Scaling em um balanceador de carga Elastic Load Balancing ajuda você a configurar uma aplicação com balanceamento de carga. O Elastic Load Balancing funciona com o Amazon EC2 Auto Scaling para distribuir o tráfego de entrada entre suas instâncias íntegras

do Amazon EC2. Isso aumenta a escalabilidade e a disponibilidade da sua aplicação. É possível habilitar o Elastic Load Balancing em várias zonas de disponibilidade para aumentar a tolerância a falhas das aplicações.

Neste tutorial, abordamos as etapas básicas para a configuração de uma aplicação com balanceamento de carga quando o grupo do Auto Scaling é criado. Quando terminar, sua arquitetura será semelhante ao diagrama a seguir:



O Elastic Load Balancing oferece suporte para diferentes tipos de balanceadores de carga. Recomendamos que você use um Application Load Balancer para este tutorial.

Para obter mais informações sobre como introduzir um balanceador de carga em sua arquitetura, consulte Use o Elastic Load Balancing para distribuir automaticamente o tráfego de entrada de aplicações no seu grupo do Auto Scaling..

Tarefas

- Pré-requisitos
- Etapa 1: Configurar um modelo de execução ou uma configuração de execução
- Etapa 2: Criar um grupo do Auto Scaling
- Etapa 3: Verificar se o balanceador de carga está anexado
- Etapa 4: Próximas etapas
- Etapa 5: Limpar
- Recursos relacionados

Pré-requisitos

- Um balanceador de carga e grupo de destino. Certifique-se de escolher as mesmas zonas de disponibilidade para o balanceador de carga que você planeja usar em seu grupo do Auto Scaling. Para obter mais informações, consulte Conceitos básicos do Elastic Load Balancing no Manual do usuário do Elastic Load Balancing.
- Um grupo de segurança para o modelo de execução ou configuração de execução. O grupo de segurança deve permitir o acesso do balanceador de carga na porta do listener (geralmente na porta 80 para tráfego HTTP) e na porta que você deseja que o Elastic Load Balancing use para verificações de integridade. Para obter mais informações, consulte a documentação aplicável:
  - Grupos de segurança de destino no Manual do usuário para Application Load Balancers
  - Grupos de segurança de destino no Manual do usuário para Network Load Balancers

  Opcionalmente, se as instâncias tiverem endereços IP públicos, também será possível permitir tráfego de SSH para conexão com as instâncias.
- (Opcional) Uma função do IAM que concede acesso ao seu aplicativo AWS a.
- (Opcional) Uma imagem de máquina da Amazon (AMI) definida como sendo o modelo de origem para suas instâncias do Amazon EC2. Para criar um agora, execute uma instância. Especifique o perfil do IAM (se tiver criado um) e os scripts de configuração de que você precisa como dados do usuário. Conecte-se à instância e personalize-a. Por exemplo, você pode instalar softwares e aplicações, copiar dados e anexar volumes adicionais do EBS. Teste suas aplicações na sua instância para garantir que ela esteja configurada corretamente. Salve esta configuração atualizada como uma AMI personalizada. Será possível terminar a instância se ela não for necessária posteriormente. Entre as instâncias executadas nessa AMI personalizada estão as personalizações que você fez quando criou a AMI.

- Uma nuvem privada virtual (VPC). Este tutorial se refere à VPC padrão, mas é possível usar a sua própria. Nesse último caso, certifique-se de que a VPC tenha uma sub-rede mapeada para cada zona de disponibilidade da região na qual você está trabalhando. No mínimo, é necessário ter duas sub-redes públicas disponíveis para criar o balanceador de carga. Você também deve ter duas sub-redes privadas ou duas sub-redes públicas para criar seu grupo do Auto Scaling e registrá-lo no balanceador de carga.

Etapa 1: Configurar um modelo de execução ou uma configuração de execução

Use um modelo de execução ou uma configuração de execução para este tutorial.

Tópicos

- Criar ou selecionar um modelo de execução
- Criar ou selecionar uma configuração de execução

Criar ou selecionar um modelo de execução

Se você já tiver um modelo de execução que gostaria de usar, selecione-o usando o procedimento a seguir.

Para selecionar um modelo de execução existente

1. Abra a página Launch templates (Modelos de execução) do console do Amazon EC2.
2. Na barra de navegação, na parte superior da tela, escolha a região onde o balanceador de carga foi criado.
3. Selecione um modelo de execução.
4. Selecione Actions (Ações), Create Auto Scaling group (Criar grupo do Auto Scaling).

Como alternativa, use o procedimento a seguir para criar um novo modelo de execução.

Para criar um modelo de execução

1. Abra a página Launch templates (Modelos de execução) do console do Amazon EC2.
2. Na barra de navegação, na parte superior da tela, escolha a região onde o balanceador de carga foi criado.

3. Escolha Criar modelo de execução.
4. Insira um nome e forneça uma descrição para a versão inicial do modelo de execução.
5. Em Application and OS Images (Amazon Machine Image) (Imagens de aplicações e sistemas operacionais [imagem de máquina da Amazon]), escolha o ID da AMI de suas instâncias. Você pode pesquisar todas as opções disponíveis AMIs ou selecionar uma AMI na lista Recentes ou Início rápido. Se você não encontrar a AMI de que precisa, escolha Procurar mais AMIs para ver o catálogo completo de AMI.
6. Em Instance type (Tipo de instância), selecione uma configuração de hardware para as suas instâncias que seja compatível com a AMI que você especificou.
7. (Opcional) Em Key pair (login) (Par de chaves - login), digite o nome do par de chaves a ser usado quando você se conectar às suas instâncias.
8. Em Network settings (Configurações de rede), expanda Advanced network configuration (Configuração de rede avançada) e execute estas ações:
    a. Escolha Add network interface (Adicionar interface de rede) para configurar a interface de rede primária.
    b. Para Auto-assign Public IP (Autoatribuir IP público): especifique se suas instâncias recebem endereços IPv4 públicos. Por padrão, o Amazon EC2 atribui um endereço IPv4 público se a instância do EC2 for iniciada em uma sub-rede padrão ou se a instância for iniciada em uma sub-rede configurada para atribuir automaticamente um endereço público. IPv4 Como alternativa, se você não precisar se conectar às suas instâncias, escolha Disable (Desabilitar) para impedir que as instâncias do seu grupo recebam tráfego diretamente da Internet. Nesse caso, elas receberão tráfego somente do load balancer.
    c. Em Security group ID (ID do grupo de segurança), especifique um grupo de segurança para suas instâncias a partir da mesma VPC que o balanceador de carga.
    d. Em Delete on termination (Excluir ao término), escolha Yes (Sim). Isso excluirá a interface de rede quando o grupo do Auto Scaling reduzir a escala na horizontal e terminará a instância na qual a interface de rede está anexada.
9. (Opcional) Para distribuir as credenciais de forma segura para as suas instâncias, em Advanced details (Detalhes avançados), IAM instance profile (Perfil de instância do IAM), digita o nome de recurso da Amazon (ARN) do seu perfil do IAM.
10. (Opcional) Para especificar os dados do usuário ou um script de configuração para suas instâncias, copie-os em Advanced details (Detalhes avançados), User data (Dados do usuário).
11. Escolha Criar modelo de execução.

12. Na página de confirmação, escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Criar ou selecionar uma configuração de execução

> **Note**
>
> Não recomendamos o uso de configurações de execução em novas aplicações porque esse é um atributo legado sem investimento planejado. Além disso, quaisquer novas contas criadas a partir de 1º de junho de 2023 não terão a opção de criar novas configurações de inicialização por meio do console. Para obter mais informações, consulte Configurações de execução do Auto Scaling.

Para selecionar uma configuração de ativação existente

1. Abra a página Launch configurations (Configurações de execução) do console do Amazon EC2.
2. Na barra de navegação superior, escolha a região onde o balanceador de carga foi criado.
3. Selecione uma configuração de ativação.
4. Selecione Actions (Ações), Create Auto Scaling group (Criar grupo do Auto Scaling).

Como alternativa, para criar uma nova configuração de ativação, use o procedimento a seguir.

Para criar uma configuração de execução

1. Abra a página Launch configurations (Configurações de execução) do console do Amazon EC2. Quando a confirmação for solicitada, escolha Exibir configurações de inicialização para confirmar que deseja visualizar a página Configurações de inicialização.
2. Na barra de navegação superior, escolha a região onde o balanceador de carga foi criado.
3. SelecioneCreate launch configuration (Criar uma configuração de execução), e insira um nome para sua configuração de execução.
4. Em Amazon machine image (AMI) (Imagem de máquina da Amazon), insira o ID da AMI para suas instâncias como critério de pesquisa.
5. Em Instance type (Tipo de instância), selecione uma configuração de hardware para sua instância.
6. Em Additional configuration (Configuração adicional), preste atenção aos seguintes campos:

a. (Opcional) Para distribuir credenciais com segurança para sua instância do EC2, em IAM instance profile (Perfil da instância do IAM), escolha seu perfil do IAM. Para obter mais informações, consulte Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2.

b. (Opcional) Para especificar os dados do usuário ou um script de configuração para sua instância, copie-os em Advanced details (Detalhes avançados), User data (Dados do usuário).

c. (Opcional) Em Advanced details (Detalhes avançados), IP address type (Tipo de endereço IP), mantenha o valor padrão. Ao criar seu grupo do Auto Scaling, é possível atribuir um endereço IP público a instâncias no seu grupo do Auto Scaling usando sub-redes que têm o atributo de endereçamento IP público habilitado, como as sub-redes padrão na VPC padrão. Como alternativa, se você não precisar se conectar às suas instâncias, escolha Do not assign a public IP address to any instances (Não atribuir um endereço IP público a nenhuma instância) para impedir que as instâncias no seu grupo recebam tráfego diretamente da Internet. Nesse caso, elas receberão tráfego somente do load balancer.

7. Em Security groups (Grupos de segurança), escolha um grupo de segurança existente na mesma VPC que o balanceador de carga. Se você mantiver Create a new security group (Criar um novo grupo de segurança) selecionado, uma regra SSH padrão será configurada para instâncias do Amazon EC2 que executem Linux. Uma função do RDP padrão é configurada para instâncias do Amazon EC2 que executem o Windows.

8. Em Key pair (login) (Par de chaves - login), escolha uma opção em Key pair options (Opções de par de chaves).

   Se já tiver configurado um par de chaves de instância do Amazon EC2, você pode escolhê-lo aqui.

   Caso você ainda não tenha um par de chaves da instância do Amazon EC2, escolha Create a new key pair (Criar um novo par de chaves) e atribua a ele um nome reconhecível. Escolha Download key pair (Fazer download do par de chaves) para fazer baixar o par de chaves para seu computador.

   > **Important**
   >
   > Não escolha Proceed without a key pair (Continuar sem um par de chaves) se você precisar se conectar à sua instância.

9. Selecione a caixa de confirmação e escolha Criar configuração de execução.
10. Marque a caixa de seleção ao lado do nome da nova configuração de execução e escolha Actions (Ações),Create Auto Scaling group (Criar grupo do Auto Scaling).

Etapa 2: Criar um grupo do Auto Scaling

Use o procedimento a seguir para continuar de onde parou depois que selecionar ou criar seu modelo de execução ou sua configuração de execução.

Para criar um grupo do Auto Scaling

1. Na página Choose launch template or configuration (Escolher o modelo ou a configuração de execução), em Auto Scaling group name (Nome do grupo do Auto Scaling) insira um nome para o grupo do Auto Scaling.
2. [Modelo de execução somente] Em Launch template (Modelo de execução), indique se o grupo do Auto Scaling usará a versão padrão, a mais recente ou uma versão específica do modelo de execução no aumento da escala na horizontal.
3. Escolha Próximo.

   A página Choose instance launch options (Escolher as opções de execução da instância) será exibida, permitindo a você escolher as configurações de rede VPC que você deseja que o grupo do Auto Scaling use e oferecendo opções de execução para instâncias spot e sob demanda (se você escolher um modelo de execução).
4. Na seção Network (Rede), para VPC, selecione a VPC usada para o balanceador de carga. Se você escolher a VPC padrão, ela será configurada automaticamente para fornecer conectividade com a Internet às instâncias. Essa VPC inclui uma sub-rede pública em cada zona de disponibilidade na região.
5. Para Availability Zones and subnets (Zonas de disponibilidade e sub-redes), selecione uma ou mais sub-redes de cada zona de disponibilidade que você deseja incluir, baseando-se em quais zonas de disponibilidade o balanceador de carga se encontra. Para obter mais informações, consulte Considerações sobre a escolha de sub-redes da VPC.
6. [Apenas modelo de execução] Na seção Instance type requirements (Requisitos de tipo de instância), use a configuração padrão para simplificar esta etapa. (Não substitua o modelo de execução.) Neste tutorial, você fará o execução apenas das Instâncias sob demanda usando o tipo de instância especificado no modelo de execução.

7. Selecione Next (Próximo) para ir até a página Configure advanced options (Configurar opções avançadas).
8. Para anexar o grupo a um balanceador de carga existente, na seção Load balancing (Balanceamento de carga), selecione Attach to an existing load balancer (Anexar a um balanceador de carga existente). É possível selecionar Choose from your load balancer target groups (Escolher entre seus grupos de destino do balanceador de carga) ou Choose from Classic Load Balancers (Escolher entre balanceadores de carga clássicos). Em seguida, você pode escolher o nome de um grupo de destino para o Application Load Balancer ou Network Load Balancer criado ou escolher o nome de um Classic Load Balancer.
9. (Opcional) Para verificações de integridade e tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Elastic Load Balancing.
10. Opcional em Tempo de carência da verificação de integridade, insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.
11. Quando terminar de configurar o grupo do Auto Scaling, escolha Skip to review (Pular para revisão).
12. Na página Review (Revisar), examine os detalhes de seu grupo do Auto Scaling. Você pode escolher Editar para fazer alterações. Ao concluir, escolha Create group (Criar grupo).

Depois de criar o grupo do Auto Scaling com o balanceador de carga anexado, o balanceador de carga registrará automaticamente as novas instâncias à medida que ficarem online. Você tem somente uma instância no momento, então não há muito para registrar. No entanto, é possível adicionar outras instâncias atualizando a capacidade desejada do grupo. Para step-by-step obter instruções, consulteAlterar a capacidade desejada de seu grupo do Auto Scaling.

Etapa 3: Verificar se o balanceador de carga está anexado

Como verificar se o balanceador de carga está associado

1. Na Auto Scaling groups page (Página Grupos do Auto Scaling) do console do Amazon EC2, marque a caixa de seleção ao lado do seu grupo do Auto Scaling.
2. Na guia Details (Detalhes), Load balancing (Balanceamento de carga) mostra os grupo de destino do balanceador de carga ou os Classic Load Balancers anexados.

3. Na guia Activity (Atividades), em Activity history (Histórico de atividades), é possível verificar se as instâncias foram executadas com êxito. A coluna Status mostra se seu grupo do Auto Scaling tem instâncias executadas com êxito. Se as instâncias não foram executadas, será possível encontrar ideias de solução de problemas para problemas comuns de execução de instâncias na Solucionar problemas do Amazon EC2 Auto Scaling.
4. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), é possível verificar se as suas instâncias estão prontas para receber tráfego. Inicialmente, suas instâncias estão no estado `Pending`. Quando uma instância está pronta para receber tráfego, seu estado é `InService`. A coluna Health Status (Status de integridade) mostra o resultado das verificações de integridade do Amazon EC2 Auto Scaling em suas instâncias. Embora uma instância possa ser marcada como íntegra, o balanceador de carga só enviará tráfego para instâncias que passarem nas verificações de integridade do balanceador de carga.
5. Verifique se suas instâncias estão registradas do balanceador de carga. Abra a página Target groups (Grupos de destino) do console do Amazon EC2. Selecione o grupo de destino e escolha a guia Destinos. Se o estado das suas instâncias for `initial`, é provavelmente porque elas ainda estão em processo de registro, ou ainda estão sendo submetidas a verificações de integridade. Quando o estado das suas instâncias for `healthy`, elas estarão prontas para uso.

Etapa 4: Próximas etapas

Agora que você concluiu este tutorial, é possível aprender mais:

- O Amazon EC2 Auto Scaling determina se a instância está íntegra com base no status das verificações de integridade que o grupo do Auto Scaling usa. Se você habilitar as verificações de integridade do balanceador de carga e houver falha nas verificações de integridade de uma instância, o grupo do Auto Scaling considerará a instância como não íntegra e a substituirá. Para obter mais informações, consulte Verificações de integridade.
- É possível expandir sua aplicação para uma zona de disponibilidade adicional na mesma região para aumentar a tolerância a falhas em caso de interrupção do serviço. Para obter mais informações, consulte Adicione uma Zona de disponibilidade.
- É possível configurar o grupo do Auto Scaling para usar uma política de escalabilidade com monitoramento do objetivo. Isso aumenta ou diminui automaticamente o número de instâncias à medida que a demanda nas instâncias for alterada. Isso permite que o grupo lide com alterações na quantidade de tráfego que a aplicação recebe. Para obter mais informações, consulte Políticas de escalabilidade de rastreamento de destino.

Etapa 5: Limpar

Após concluir os recursos que você criou para este tutorial, será necessário considerar limpá-los para evitar cobranças desnecessárias.

Para excluir seu grupo do Auto Scaling

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.
3. Escolha Excluir.
4. Quando a confirmação for solicitada, digite **`delete`** para confirmar a exclusão do grupo do Auto Scaling especificado e, em seguida, escolha Excluir.

   Um ícone de carregamento na coluna Name (Nome) indica que o grupo do Auto Scaling está sendo excluído. Quando a exclusão tiver ocorrido, as colunas Desejado, Mínimo e Máximo exibirão `0` instâncias para o grupo do Auto Scaling. São necessários alguns minutos para encerrar a instância e excluir o grupo. Atualize a lista para ver o estado atual.

Ignore esse procedimento se quiser manter seu modelo de execução.

Para excluir seu modelo de execução

1. Abra a página Launch templates (Modelos de execução) do console do Amazon EC2.
2. Selecione seu modelo de execução.
3. Escolha Actions (Ações), Delete template (Excluir modelo).
4. Quando a confirmação for solicitada, digite **`Delete`** para confirmar a exclusão do modelo de execução especificado e, em seguida, escolha Excluir.

Ignore esse procedimento se quiser manter sua configuração de execução.

Para excluir sua configuração de ativação

1. Abra a página Launch configurations (Configurações de execução) do console do Amazon EC2.
2. Selecione sua configuração de execução.
3. Escolha Ações, Excluir configuração de execução.
4. Quando a confirmação for solicitada, escolha Excluir.

Ignore o procedimento a seguir se desejar manter o balanceador de carga para uso futuro.

Para excluir o balanceador de carga

1. Abra a página Load balancers (Balanceadores de carga) do console do Amazon EC2.
2. Selecione o balanceador de carga e Actions (Ações), Delete (Excluir).
3. Quando a confirmação for solicitada, escolha Sim, excluir.

Para excluir seu grupo de destino

1. Abra a página Target groups (Grupos de destino) do console do Amazon EC2.
2. Selecione o grupo de destino e escolha Actions (Ações), Delete (Excluir).
3. Quando a confirmação for solicitada, escolha Sim, excluir.

Recursos relacionados

Com CloudFormation, você pode criar e provisionar implantações de AWS infraestrutura de forma previsível e repetida, usando arquivos de modelo para criar e excluir uma coleção de recursos juntos como uma única unidade (uma pilha). Para obter mais informações, consulte o Guia do usuário AWS CloudFormation.

Para obter um passo a passo que usa um modelo de pilha para provisionar um grupo do Auto Scaling e um Application Load Balancer, consulte Passo a passo: Criar uma aplicação dimensionado e com balanceamento de carga no AWS CloudFormation Guia do usuário. Use o passo a passo e o modelo de amostra como ponto de partida para criar modelos semelhantes que atendam às suas necessidades.

Execução do Auto Scaling e modelos de execução

Um modelo de execução é semelhante a uma configuração de execução, uma vez que especifica informações de configuração de instância. Isso inclui o ID da Imagem de máquina da Amazon (AMI), o tipo de instância, um par de chaves, grupos de segurança e outros parâmetros que você usa para iniciar instâncias do EC2. No entanto, definir um modelo de execução em vez de uma configuração de execução permite ter várias versões de um modelo de execução.

Com o versionamento dos modelos de execução, você pode criar um subconjunto do conjunto completo de parâmetros. Em seguida, você pode reutilizá-lo para criar outras versões do mesmo modelo de execução. Por exemplo, você pode criar um modelo de execução que defina uma configuração base sem uma AMI ou um script de dados do usuário. Depois de criar o modelo de execução, você pode criar uma nova versão e adicionar a AMI e os dados do usuário que têm a versão mais recente da aplicação para teste. Isso resulta em duas versões do modelo de execução. Armazenar uma configuração base ajuda você a manter os parâmetros de configuração geral necessários. Você pode criar uma nova versão do modelo de execução da configuração base sempre que quiser. Você também pode excluir as versões usadas para testar sua aplicação quando não precisar mais delas.

Recomendamos que você use modelos de execução para garantir que esteja acessando os recursos e melhorias mais recentes. Nem todos os recursos do Amazon EC2 Auto Scaling estão disponíveis quando você usa configurações de execução. Por exemplo, não é possível criar um grupo do Auto Scaling que execute instâncias spot e sob demanda ou que especifique vários tipos de instância. É necessário usar um modelo de execução para configurar esses recursos. Para obter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

Com modelos de execução, você também pode usar recursos mais recentes do Amazon EC2. Isso inclui parâmetros do Systems Manager (ID da AMI), a geração atual de volumes da IOPS provisionadas do EBS (io2), marcação de volumes do EBS, instâncias T2 Unlimited, reservas de capacidade Capacity Blocks e hosts dedicados, para citar alguns.

Ao criar um modelo de execução, todos os parâmetros são opcionais. No entanto, se um modelo de execução não especificar uma AMI, você não poderá adicionar a AMI ao criar seu grupo do Auto Scaling. Se você especificar uma AMI, mas nenhum tipo de instância, poderá adicionar um ou mais tipos de instância ao criar seu grupo do Auto Scaling.

Conteúdo

- Permissões para trabalhar com modelos de inicialização

- Operações de API compatíveis com os modelos de execução
- Criar um modelo de execução para um grupo do Auto Scaling
- Criar um modelo de execução usando configurações avançadas
- Migre seus grupos do Auto Scaling para modelos de execução
- Migre AWS CloudFormation pilhas para modelos de lançamento
- Exemplos para criar e gerenciar modelos de lançamento com o AWS CLI
- Use AWS Systems Manager parâmetros em vez de AMI IDs em modelos de lançamento

Permissões para trabalhar com modelos de inicialização

Os procedimentos nesta seção pressupõem que você já tenha as permissões necessárias para criar modelos de execução. Para obter informações sobre como um administrador concede permissões a você, consulte Controlar o acesso aos modelos de execução com permissões do IAM no Guia do usuário do Amazon EC2 .

Observe que se você não tiver permissões suficientes para usar e criar recursos especificados em um modelo de execução, você receberá um erro informando que não está autorizado a usar o modelo de execução ao tentar especificá-lo para um grupo do Auto Scaling. Para obter mais informações, consulte Solucionar problemas do Amazon EC2 Auto Scaling: modelos de execução.

Para obter exemplos de políticas do IAM que permitem chamar as operações da API `CreateAutoScalingGroup`, `UpdateAutoScalingGroup` e `RunInstances` com um modelo de inicialização, consulte Controle o uso do modelo de execução do Amazon EC2 em grupos do Auto Scaling.

Operações de API compatíveis com os modelos de execução

Para obter uma lista de operações de API suportadas por modelos de execução, consulte Ações do Amazon EC2 na Referência de API do Amazon EC2.

Criar um modelo de execução para um grupo do Auto Scaling

Antes de criar um grupo do Auto Scaling usando um modelo de execução, será necessário criar um modelo de execução que contenha as informações de configuração para iniciar uma instância, incluindo o ID da Imagem de máquina da Amazon (AMI).

Siga os procedimentos abaixo para criar novos modelos de execução.

Conteúdo

- Criar seu modelo de execução (console)
- Alterar as configurações da interface de rede padrão (console)
- Modificar a configuração do armazenamento (console)
- Criar um modelo de execução com base em uma instância existente (console)
- Recursos relacionados
- Limitações

> **Important**
>
> Os parâmetros do modelo de execução não são totalmente validados quando ele é criado. Se você especificar valores incorretos para parâmetros, ou se não usar combinações de parâmetro compatíveis, nenhuma instância poderá ser iniciada usando esse modelo de execução. Certifique-se de especificar os valores corretos para os parâmetros e use as combinações de parâmetros com suporte. Por exemplo, para iniciar instâncias com uma AMI AWS Graviton ou Graviton2 baseada em Arm, é necessário especificar um tipo de instância compatível com Arm. Para obter mais informações, consulte Restrições de um modelo de execução na Guia do usuário do Amazon EC2.

Criar seu modelo de execução (console)

As etapas a seguir descrevem como configurar um modelo básico de execução:

- Especificar a imagem de máquina da Amazon (AMI) da qual as instâncias serão iniciadas.
- Escolher um tipo de instância compatível com a AMI que você especificar.
- Especificar o par de chaves a ser usado ao conectar-se a instâncias, por exemplo, usando SSH.
- Adicionar um ou mais grupos de segurança para permitir acesso relevante às instâncias de uma rede externa.
- Especifique se deseja adicionar volumes adicionais a cada instância.
- Adicionar tags personalizadas (pares chave-valor) às instâncias e aos volumes.

Para criar um modelo de execução

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.

2. No painel de navegação, escolha Instances e, em seguida, Launch Templates.
3. Escolha Criar modelo de execução. Insira um nome e forneça uma descrição para a versão inicial do modelo de execução.
4. (Opcional) Em Orientação do ajuste de escala automático, marque a caixa de seleção para que o Amazon EC2 forneça orientações para ajudar a criar um modelo para uso com o Amazon EC2 Auto Scaling.
5. Em Launch template contents (Conteúdo do modelo de execução), preencha todos os campos obrigatórios e campos opcionais, conforme necessário.

   a. Imagens de aplicações e sistemas operacionais (Amazon Machine Image): (Obrigatório) escolha o ID da AMI para suas instâncias. Você pode pesquisar todas as opções disponíveis AMIs ou selecionar uma AMI na lista Recentes ou Início rápido. Se você não encontrar a AMI de que precisa, escolha Procurar mais AMIs para ver o catálogo completo de AMI.

      Para escolher uma AMI personalizada, primeiro é necessário criar uma AMI desde uma instância personalizada. Para mais informações, consulte Criar uma AMI com suporte do Amazon EBS na Guia do Usuário do Amazon EC2.

   b. Em Instance type (Tipo de instância), escolha um único tipo de instância compatível com a AMI que você especificou.

      Como alternativa, para usar a seleção de tipo de instância baseada em atributos, escolha Avançado, Especificar atributos de tipo de instância e, em seguida, especifique as seguintes opções:

      - Número de v CPUs: insira o número mínimo e máximo de CPUs v. Para indicar que não há limites, insira um mínimo de 0 e mantenha o máximo em branco.
      - Amount of memory (MiB) (Quantidade de memória): insira a quantidade mínima e máxima de memória, em MiB. Para indicar que não há limites, insira um mínimo de 0 e mantenha o máximo em branco.
      - Expanda Optional instance type attributes (Atributos de tipo de instância opcionais) e escolha Add attribute (Adicionar atributo) para limitar ainda mais os tipos de instâncias que podem ser usadas para atender à capacidade desejada. Para obter informações sobre cada atributo, consulte InstanceRequirementsRequesta Referência de API do Amazon EC2.

- Tipos de instância resultantes: você pode visualizar os tipos de instância que correspondem aos requisitos de computação especificados, como vCPUs, memória e armazenamento.
- Para excluir tipos de instância, escolha Add attribute (Adicionar atributo). Do Attribute list (lista de Atribuição), escolha Excluded instances types (Tipos de instâncias excluídas). Na lista Attribute value (Valor do atributo), selecione os tipos de instância a serem excluídos.

c. Key pair (login) (Par de chaves): para Key pair name (Nome do par de chaves), escolha um par de chaves existente ou escolha Create new key pair (Criar um novo par de chaves) para criar um novo. Para obter mais informações, consulte Pares de chaves do Amazon EC2 no Guia do usuário do Amazon EC2 .

d. Network settings (Configurações de rede): para Firewall (security groups) (grupos de segurança) ou deixe em branco e configure um ou mais grupos de segurança como parte da interface de rede. Para obter mais informações, consulte Grupos de segurança do Amazon EC2 para instâncias do Linux no Guia do usuário do Amazon EC2.

   Se você não especificar nenhum grupo de segurança em seu modelo de execução, o Amazon EC2 usará o grupo de segurança padrão para a VPC na qual seu grupo do Auto Scaling iniciará instâncias. Por padrão, esse grupo de segurança não permite tráfego de entrada de redes externas. Para obter mais informações, consulte Grupos de segurança padrão para você VPCs no Guia do usuário da Amazon VPC.

e. Execute um destes procedimentos:

   - Altere as configurações da interface de rede padrão Por exemplo, você pode ativar ou desativar o recurso de IPv4 endereçamento público, que substitui a configuração de atribuição automática de IPv4 endereços públicos na sub-rede. Para obter mais informações, consulte Alterar as configurações da interface de rede padrão (console).
   - Ignore essa etapa para manter as configurações da interface de rede padrão.

f. Execute um destes procedimentos:

   - Modificar a configuração do armazenamento Para obter mais informações, consulte Modificar a configuração do armazenamento (console).
   - Ignore essa etapa para manter a configuração de armazenamento padrão.

g. Em Resource tags (Etiquetas de recurso), especifique as etiquetas fornecendo combinações de chave e valor. Se você especificar tags de instância em seu modelo de execução e optar por propagar tags de seu grupo do Auto Scaling para suas instâncias, todas as tags serão mescladas. Se a mesma chave da etiqueta for especificada para uma

etiqueta no modelo de execução e uma etiqueta no grupo do Auto Scaling, então, o valor da etiqueta do grupo terá precedência.

6. Definir configurações avançadas (opcional) Por exemplo, você pode escolher um perfil do IAM que sua aplicação possa usar ao acessar outros recursos da AWS , ou especificar os dados do usuário da instância que podem ser usados para executar tarefas de configuração automatizadas comuns após o início de uma instância. Para obter mais informações, consulte Criar um modelo de execução usando configurações avançadas.
7. Quando você estiver pronto para criar seu modelo de execução, escolha Create launch template (Criar modelo de execução).
8. Para criar um grupo do Auto Scaling, escolha Create Auto Scaling group (Criar grupo do Auto Scaling) na página de confirmação.

Alterar as configurações da interface de rede padrão (console)

As interfaces de rede fornecem conectividade com outros recursos em sua VPC e na Internet. Para obter mais informações, consulte Fornecer conectividade de rede para suas instâncias do Auto Scaling usando a Amazon VPC.

Esta seção mostra como alterar as configurações padrão da interface de rede. Por exemplo, você pode definir se deseja atribuir um IPv4 endereço público a cada instância em vez de usar como padrão a configuração de atribuição automática de IPv4 endereços públicos na sub-rede.

Considerações e limitações

Ao alterar as configurações padrão da interface de rede, lembre-se das seguintes considerações e limitações:

- Será necessário configurar o grupo de segurança como parte da interface de rede, e não na seção Grupos de segurança do modelo. Não é possível especificar grupos de segurança nos dois locais.
- Se você especificar um ID de interface de rede existente, poderá iniciar apenas uma instância. Para fazer isso, você deve usar o AWS CLI ou um SDK para criar o grupo Auto Scaling. Ao criar o grupo, é necessário especificar a zona de disponibilidade, mas não o ID da sub-rede. Além disso, você pode especificar uma interface de rede existente somente se ela tiver um índice de dispositivo de 0.
- Você não pode atribuir automaticamente um IPv4 endereço público se especificar mais de uma interface de rede. Você também não pode especificar índices de dispositivos duplicados em interfaces de rede. As interfaces de rede primária e secundária residem na mesma sub-rede.

- Quando uma instância é iniciada, é atribuído um endereço privado automaticamente para cada interface de rede. O endereço vem do intervalo CIDR da sub-rede na qual a instância é iniciada. Para obter informações sobre como especificar blocos CIDR (ou intervalos de endereços IP) para sua VPC ou sub-rede, consulte o Manual do usuário da Amazon VPC.

Para alterar as configurações da interface de rede padrão

1. Em Network settings (configurações de rede), expanda Advanced network configuration (configuração de rede avançada).
2. Escolha Add network interface (Adicionar interface de rede) para configurar a interface de rede primária, prestando atenção aos seguintes campos:

   a. Device index (Índice do dispositivo): mantenha o valor padrão, 0, para aplicar suas alterações à interface de rede primária (eth0).

   b. Interface de rede: mantenha o valor padrão, New interface (Nova interface), para que o Amazon EC2 Auto Scaling crie automaticamente uma nova interface de rede quando uma instância for iniciada. Como alternativa, você pode escolher uma interface de rede existente e disponível com um índice de dispositivo de 0, mas isso limita seu grupo do Auto Scaling a uma instância.

   c. Description (Descrição): insira um nome descritivo.

   d. Subnet (Sub-rede): mantenha a configuração padrão Don't include in launch template (Não incluir no modelo de inicialização).

      Se a AMI especificar uma sub-rede para a interface de rede, isso resultará em um erro. Recomendamos desligar Auto Scaling guidance (Orientação do Auto Scaling) como uma solução alternativa. Depois de fazer essa alteração, você não receberá uma mensagem de erro. No entanto, independentemente de onde a sub-rede é especificada, as configurações de sub-rede do grupo do Auto Scaling têm precedência e não podem ser substituídas.

   e. Atribuição automática de IP público: altere se sua interface de rede com um índice de dispositivo de 0 recebe um IPv4 endereço público. Por padrão, as instâncias em uma sub-rede padrão recebem um IPv4 endereço público, enquanto as instâncias em uma sub-rede não padrão não. Selecione Enable (Habilitar) ou Disable (Desabilitar) para substituir a configuração padrão da sub-rede.

   f. Security groups (Grupos de segurança): selecione um ou mais grupos de segurança para a interface de rede. Cada grupo de segurança deve ser configurado para a VPC na qual seu

grupo do Auto Scaling iniciará instâncias. Para obter mais informações, consulte Grupos de segurança do Amazon EC2 para instâncias do Linux no Guia do usuário do Amazon EC2.

g. Delete on termination (Excluir no encerramento(: escolha Yes (Sim) para excluir a interface de rede quando a instância for encerrada, ou escolha No (Não) para manter a interface de rede.

h. Elastic Fabric Adapter (Adaptador de malha elástica): para dar suporte a casos de uso de computação de alto desempenho e de machine learning, altere a interface de rede para uma interface de rede do Elastic Fabric Adapter. Para obter mais informações, consulte Elastic Fabric Adapter no Manual do usuário do Amazon EC2.

i. Network card index (Índice da placa de rede): escolha 0 para anexar a interface de rede primária à placa de rede com um índice de dispositivo de 0. Se essa opção não estiver disponível, mantenha o valor padrão, Don't include in launch template (Não incluir no modelo de inicialização). Anexar a interface de rede a uma placa de rede específica está disponível apenas para tipos de instância compatíveis. Para obter mais informações, consulte Network cards (Placas de rede) no Manual do usuário do Amazon EC2.

j. ENA Express: Para tipos de instâncias que suportam o ENA Express, selecione Habilitar para ativar o ENA Express ou Desabilitar para desativá-lo. Para obter mais informações, consulte Melhorar o desempenho de rede com o ENA Express em instâncias do Linux no Guia do usuário do Amazon EC2.

k. ENA Express UDP: Se você habilitar o ENA Express, poderá usá-lo opcionalmente para tráfego UDP. Escolha Ativar para ativar o UDP do ENA Express ou Desativar para desativá-lo.

3. Para adicionar uma interface de rede secundária, escolha Add network interface (Adicionar interface de rede).

Modificar a configuração do armazenamento (console)

Você pode modificar a configuração de armazenamento para instâncias executadas de uma AMI baseada no Amazon EBS ou de uma AMI com armazenamento de instâncias. É possível especificar volumes EBS adicionais para anexar às instâncias. A AMI inclui um ou mais volumes de armazenamento, incluindo o volume raiz (Volume 1 (AMI Root [Raiz da AMI]).

Para modificar a configuração do armazenamento

1. Em Configure storage (Configurar armazenamento), modifique o tamanho ou o tipo de volume.

Se o valor especificado para o tamanho do volume estiver fora dos limites do tipo de volume ou menor que o tamanho do snapshot, uma mensagem de erro será exibida. Para ajudar a resolver o problema, esta mensagem fornece o valor mínimo ou máximo que o campo pode aceitar.

Somente volumes associados a uma AMI baseada no Amazon EBS são exibidos. Para exibir informações sobre a configuração de armazenamento de uma instância executada a partir de uma AMI com armazenamento de instâncias, escolha Show details (Mostrar detalhes) na seção volumes de armazenamento de instância.

Para especificar todos os parâmetros de volume do EBS, alterne para a visualização Advanced (Avançada) no canto superior direito.

2. Para opções avançadas, expanda o volume que você deseja modificar e configure o volume da seguinte forma:

   a. Storage type (Tipo de armazenamento): o tipo de volume (EBS ou temporário) a ser associado à instância. O tipo de volume de armazenamento de instância (temporário) só estará disponível se você selecionar um tipo de instância compatível com ele. Para mais informações, consulte Volumes do Amazon EBS no Guia do Usuário do Amazon EBS e Armazenamento de instâncias do Amazon EC2 no Guia do Usuário do Amazon EC2.

   b. Device Name (Nome do dispositivo): selecione na lista de nomes de dispositivo disponíveis para o volume.

   c. Snapshot: selecione o snapshot do qual o volume será criado. Também é possível pesquisar snapshots públicos e compartilhados que estão disponíveis, inserindo texto no campo Snapshot.

   d. Size (GiB) (Tamanho): para volumes do EBS, especifique um tamanho de armazenamento. Se você tiver selecionado uma AMI e uma instância que estejam qualificadas para o nível gratuito, tenha em mente que para permanecer no nível gratuito, seu armazenamento total deverá ficar abaixo de 30 GiB. Para obter mais informações, consulte Restrições de tamanho e configuração de um volume do EBS no Guia do usuário do Amazon EBS.

   e. Volume type (Tipo de volume): para volumes do EBS, escolha o tipo de volume. Para obter mais informações, consulte Tipos de volumes do Amazon EBS no Guia do usuário do Amazon EC2.

   f. IOPS: se você selecionou um tipo de volume SSD de IOPS provisionada (`io1`e`io2`) ou SSD de uso geral (`gp3`), poderá inserir o número de I/O operações por segundo (IOPS) que o volume pode suportar. Isso é necessário para volumes io1, io2 e gp3. Isso não é compatível com volumes gp2, st1, sc1 ou volumes padrão.

g. Delete on termination (Excluir ao término): em volumes do EBS, escolha Yes (Sim), para excluir o volume quando a instância associada for terminada, ou escolha No (Não) para manter o volume.

h. Encrypted: (Criptografado): se o tipo de instância oferecer suporte à criptografia do EBS, será possível escolher Yes (Sim) para habilitar criptografia para o volume. Se você tiver habilitado a criptografia por padrão nessa região, a criptografia estará habilitada para você. Para obter mais informações, consulte Criptografia do Amazon EBS e Ativação da criptografia do Amazon EBS por padrão no Guia do usuário do Amazon EBS.

   O efeito padrão obtido na configuração desse parâmetro varia de acordo com a opção de origem de volume, conforme descrito na tabela a seguir. Em todos os casos, você deve ter permissão para usar o especificado AWS KMS key.

   Resultados da criptografia

Se o parâmetro **`Encrypted`** estiver definido como...	E se a origem de volume for...	O estado de criptografia padrão será...	Observações
Não	Novo volume (vazio)	Não criptografado*	N/D
	Snapshot não criptografado pertencente a você	Não criptografado*	
	Snapshot criptografado pertencente a você	Criptografado pela mesma chave	
	Snapshot não criptografado compartilhado com você	Não criptografado*	
	Snapshot criptografado compartilhado com você	Criptografado pela chave do KMS padrão	

Se o parâmetro **`Encrypted`** estiver definido como...	E se a origem de volume for...	O estado de criptografia padrão será...	Observações
Sim	Novo volume	Criptografado pela chave do KMS padrão	Para usar uma chave KMS não padrão, especifique um valor para o parâmetro de chave KMS key.
	Snapshot não criptografado pertencente a você	Criptografado pela chave do KMS padrão	
	Snapshot criptografado pertencente a você	Criptografado pela mesma chave	
	Snapshot não criptografado compartilhado com você	Criptografado pela chave do KMS padrão	
	Snapshot criptografado compartilhado com você	Criptografado pela chave do KMS padrão	

* Se encryption by default (criptografia por padrão) estiver habilitado, todos os volumes recém-criados (estando ou não o parâmetro Encrypted (Criptografado) definido como Yes (Sim)) serão criptografados usando a chave KMS padrão. Se definir ambos os parâmetros Encrypted (Criptografado) e Key KMS (Chave), então permite especificar uma chave KMS não padrão.

i. KMS Key (Chave do KMS): se você escolheuYes (Sim) para Encrypted (Criptografado), deve selecionar uma chave gerenciada pelo cliente a ser usada para criptografar o volume. Se tiver habilitado a criptografia por padrão nessa região, a chave gerenciada pelo cliente padrão será selecionada para você. Você pode selecionar uma chave diferente ou especificar o ARN de qualquer chave gerenciada pelo cliente que você criou anteriormente usando o AWS Key Management Service.

3. Para especificar volumes adicionais a serem anexados às instâncias executadas por esse modelo de execução, escolha Add new volume (Adicionar novo volume).

Criar um modelo de execução com base em uma instância existente (console)

Para criar um modelo de execução a partir de uma instância existente

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, em Instances (Instâncias), escolha Instances (Instâncias).
3. Selecione a instância e escolha Actions (Ações), Image and templates (Imagem e modelos), Create template from instance (Criar modelo a partir da instância).
4. Forneça um nome e a uma descrição.
5. Em Auto Scaling guidance (Guia do Auto Scaling), marque a caixa de seleção.
6. Ajuste todas as configurações necessárias, e escolha Create launch template (Criar modelo de execução).
7. Para criar um grupo do Auto Scaling, escolha Create Auto Scaling group (Criar grupo do Auto Scaling) na página de confirmação.

Recursos relacionados

Fornecemos alguns trechos de modelos JSON e YAML que você pode usar para entender como declarar modelos de lançamento em seus modelos de pilha. CloudFormation Para obter mais informações, consulte AWS::EC2::LaunchTemplateas CloudFormation seções Criar modelos de lançamento com o Guia AWS CloudFormation do usuário.

Para obter mais informações sobre como criar um modelo de execução, consulte Iniciar uma instância com um modelo de execução no Guia do usuário do Amazon EC2.

Limitações

- Embora você possa especificar uma sub-rede em um modelo de execução, isso não é necessário se você usar o modelo de execução somente para criar grupos do Auto Scaling. Você não pode especificar a sub-rede para um grupo do Auto Scaling especificando a sub-rede em um modelo de execução. As sub-redes do grupo do Auto Scaling são retiradas da própria definição de recursos do grupo do Auto Scaling.

- Sobre outras limitações em interfaces de rede definidas pelo usuário, consulte Alterar as configurações da interface de rede padrão (console).

Criar um modelo de execução usando configurações avançadas

Este tópico descreve como criar um modelo de inicialização com configurações avançadas a partir do Console de gerenciamento da AWS.

Para criar uma planilha de execução usando configurações avançadas

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, em Instâncias, selecione Modelos de Inicialização e, em seguida, escolha Criar modelo de inicialização.
3. Configure o seu modelo de inicialização conforme descrito nos seguintes tópicos:
   - Configurações necessárias
   - Configurações avançadas
4. Escolha Criar modelo de execução.

Configurações necessárias

Ao criar uma planilha de inicialização, é necessário incluir as configurações necessárias a seguir.

Nome do modelo de execução

Digite um nome exclusivo que descreva o modelo de inicialização.

Imagens de aplicações e sistemas operacionais (imagem de máquina da Amazon)

Escolha a imagem de máquina da Amazon (AMI) que você deseja usar. Você pode pesquisar ou procurar a AMI que deseja usar. Para obter a melhor eficiência de escala, escolha uma AMI personalizada que esteja totalmente configurada para iniciar uma instância com o código da sua aplicação e que exija poucas modificações na inicialização.

Tipo de instância

Escolha um tipo de instância compatível com a AMI. Você pode ignorar a adição de um tipo de instância ao seu modelo de inicialização se planeja usar vários tipos de instâncias incorporados

na definição de recurso do próprio grupo do Auto Scaling. Um tipo de instância só é necessário se você não planeja criar um grupo de instâncias mistas.

Configurações avançadas

As configurações avançadas são opcionais. Se você não definir nenhuma configuração avançada, os recursos específicos não serão adicionados às suas instâncias.

Amplie a seção Detalhes avançados para ver as configurações avançadas. As seções a seguir descrevem as configurações avançadas mais úteis em que é necessário se concentrar ao criar um modelo de inicialização para um grupo de Auto Scaling. Para obter mais informações, consulte Detalhes Avançados na Guia do Usuário do Amazon EC2.

Perfil de instância do IAM

O perfil da instância contém o perfil do IAM que você deseja utilizar. Quando seu grupo do Auto Scaling inicia uma instância do EC2, as permissões definidas no perfil do IAM associada são concedidas às aplicações em execução na instância. Para obter mais informações, consulte Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2.

Termination protection

Quando ativado, esse atributo impede que os usuários terminem uma instância usando o console do Amazon EC2, comandos da CLI e operações de API. A proteção contra terminação oferece uma proteção extra contra terminação acidental. Isso não impede que o Amazon EC2 Auto Scaling encerre uma instância. Para controlar quais instâncias o Amazon EC2 Auto Scaling pode terminar, consulte Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância.

CloudWatch Monitoramento detalhado

Você pode ativar o monitoramento detalhado de suas instâncias do EC2 para permitir que elas enviem dados métricos para a Amazon CloudWatch em intervalos de 1 minuto. Por padrão, as instâncias do EC2 enviam dados métricos em intervalos CloudWatch de 5 minutos. Aplicam-se cobranças adicionais do . Para obter mais informações, consulte Configurar monitoramento para instâncias do Auto Scaling.

Especificação de crédito

O Amazon EC2 oferece instâncias de desempenho expansível, como T2, T3 e T3a, que permitem que as aplicações ultrapassem o desempenho de linha de base da CPU quando necessário. Por

padrão, essas instâncias podem ser expandidas por um tempo limitado antes que a utilização da CPU seja controlada. Opcionalmente, você pode ativar o modo ilimitado para que as instâncias possam ser expandidas além da linha de base pelo tempo que for necessário. Isso permite que as aplicações mantenham o alto desempenho da CPU quando necessário. Taxas adicionais podem ser aplicadas. Para mais informações, consulte Use um grupo de Auto Scaling para iniciar uma instância de desempenho escalável como Ilimitada no Guia do Usuário do Amazon EC2.

Nome do grupo de posicionamento

Você pode especificar um grupo de posicionamento e usar um cluster ou uma estratégia de partição para influenciar a forma como suas instâncias estão fisicamente localizadas no data center da AWS . Para grupos pequenos do Auto Scaling, você também pode usar a estratégia de propagação. Para obter mais informações, consulte Grupos de posicionamento, no Guia do usuário do Amazon EC2.

Há algumas considerações ao usar grupos de posicionamento com grupos do Auto Scaling:

- Se um grupo de posicionamento for especificado no modelo de inicialização e no grupo do Auto Scaling, o grupo de posicionamento do grupo do Auto Scaling terá precedência.
- Em CloudFormation, tenha cuidado ao definir um grupo de posicionamento no modelo de lançamento. O Amazon EC2 Auto Scaling iniciará as instâncias no grupo de posicionamento especificado. No entanto, não CloudFormation receberá sinais dessas instâncias se você usar um UpdatePolicycom seu grupo de Auto Scaling (embora isso possa mudar no futuro).

Opção de compra

Você pode escolher Request Spot instances (Solicitar instâncias spot) para solicitar ao preço spot, limitado ao preço sob demanda e escolha Customize (Personalizar) para alterar as configurações padrão da instância spot. Para um grupo do Auto Scaling, é necessário especificar uma solicitação única sem data de término (o padrão). Para obter mais informações, consulte Solicitar instâncias spot para aplicações flexíveis e com tolerância a falhas. Esta configuração pode ser útil em circunstâncias especiais, mas, em geral, é melhor não especificá-la e, em seu lugar, é melhor criar um grupo misto de instâncias. Para obter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

Se você especificar uma solicitação de instância spot em seu modelo de execução, não poderá criar um grupo misto de instâncias. Se você tentar usar um modelo de execução que solicite instâncias spot com um grupo misto de instâncias, você receberá a seguinte mensagem de erro: `Incompatible launch template: You cannot use a launch template that is set to request Spot Instances (InstanceMarketOptions) when you configure`

```
an Auto Scaling group with a mixed instances policy. Add a different
launch template to the group and try again.
```

Capacity Reservation

As Reservas de Capacidade permitem reservar capacidade para suas instâncias do Amazon EC2 em uma Zona de Disponibilidade específica por qualquer duração. Consulte mais informações em On-Demand Capacity Reservations no Guia do usuário do Amazon EC2.

Você pode escolher se deseja iniciar instâncias em:

- qualquer reserva de capacidade aberta (aberta)
- uma reserva de capacidade específica (destino por ID)
- um grupo de reservas de capacidade (destino por grupo)

Para selecionar uma reserva de capacidade específica, o tipo de instância em seu modelo de inicialização deve corresponder ao tipo de instância da reserva. Quando você criar o grupo do Auto Scaling, use a mesma zona de disponibilidade que a reserva de capacidade. Dependendo do Região da AWS que você escolher, você pode escolher como alvo um Bloco de Capacidade. Para obter mais informações, consulte Use Capacity Blocks para workloads de machine learning..

Para escolher um grupo de Reservas de Capacidade, consulte Reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas. Ao selecionar um grupo de reservas de capacidade, você pode ter capacidade distribuída em várias zonas de disponibilidade para melhorar a resiliência.

Locação

O Amazon EC2 oferece três opções para a locação de suas instâncias do EC2:

- Compartilhada (Compartilhada): várias Contas da AWS podem compartilhar o mesmo hardware físico. Essa é a opção de locação padrão ao iniciar uma instância.
- Instância dedicada (Dedicada) – sua instância é executada em hardware de locatário único. Nenhum outro cliente da AWS compartilha o mesmo servidor físico. Para obter mais informações, consulte Instâncias dedicadas no Guia do usuário do Amazon EC2.
- Host dedicado: (Host Dedicado) um host dedicado é um servidor físico dedicado para seu uso. O uso de hosts dedicados facilita o uso da opção traga a sua própria licença (BYOL) que têm requisitos de hardware dedicados ao EC2 e atendem aos casos de uso de conformidade. Se você escolher essa opção, deverá fornecer um grupo de recursos de host para o grupo de recursos de host de locação. Para obter mais informações, consulte Hosts dedicados no Guia do usuário do Amazon EC2.

O suporte para Hosts Dedicados está disponível apenas se você especificar um grupo de recursos de host. Não é possível direcionar um ID de host específico nem usar afinidade de posicionamento de host.

- Se você tentar usar um modelo de inicialização que especifique um ID de host, receberá a seguinte mensagem de erro: `Incompatible launch template: Tenancy host ID is not supported for Auto Scaling.`
- Se você tentar usar um modelo de execução que especifique a afinidade de posicionamento do host, receberá a seguinte mensagem de erro: `Incompatible launch template: Auto Scaling does not support host placement affinity.`

Grupo de recursos de host de locação

Com AWS License Manager, você pode trazer suas próprias licenças AWS e gerenciá-las centralmente. Um grupo de recursos de host é um grupo de hosts dedicados vinculados a uma configuração de licença específica do License Manager. Os grupos de recursos de host permitem que você inicie facilmente instâncias do EC2 em hosts dedicados que atendam às suas necessidades de licenciamento de software. Você não precisa alocar manualmente os hosts dedicados com antecedência. Eles são criados automaticamente conforme necessário. Observe que, quando você associa uma AMI a uma configuração de licença, essa AMI só pode ser associada a um grupo de recursos de host por vez. Para obter mais informações, consulte Grupos de recursos de host no AWS License Manager no Guia do usuário do License Manager.

Configurações de licença

Com essa configuração, você pode especificar uma configuração de licença para suas instâncias sem restringir a locação delas a hosts dedicados. A configuração da licença rastreia as licenças de software implantadas nas instâncias para que você possa monitorar o uso e a conformidade da licença. Para mais informações, consulte Criar uma licença gerenciada pelo próprio usuário no Guia do Usuário do License Manager.

Metadados acessíveis

Você pode escolher se deseja habilitar ou desabilitar o acesso ao endpoint HTTP do serviço de metadados da instância. Por padrão, o endpoint de HTTP está habilitado. Se você optar por desabilitar o endpoint, o acesso aos metadados da instância será desativado. Você pode especificar a condição a ser exigida IMDSv2 somente quando o endpoint HTTP estiver habilitado. Consulte mais informações em Configurar as opções de metadados da instância no Guia do usuário do Amazon EC2.

Versão de metadados

Você pode optar por exigir o uso do Instance Metadata Service Version 2 (IMDSv2) ao solicitar metadados da instância. Se você não especificar um valor, o padrão é oferecer suporte a ambos IMDSv1 IMDSv2 e. Consulte mais informações em Configurar as opções de metadados da instância no Guia do usuário do Amazon EC2.

Limite de saltos da resposta do token de metadados

Você pode definir o número permitido de saltos de rede para o token de metadados. Se você não especificar um valor, o padrão será 1. Consulte mais informações em Configurar as opções de metadados da instância no Guia do usuário do Amazon EC2.

Dados do usuário

É possível personalizar e concluir a configuração das instâncias no momento da inicialização, especificando scripts de shell ou diretivas de inicialização na nuvem como dados do usuário. Os dados do usuário são executados quando a instância é iniciada, permitindo a instalação automática de aplicações, dependências ou personalizações no momento da inicialização. Para obter mais informações, consulte Executar comandos na instância do Linux no lançamento no Guia do usuário do Amazon EC2.

Se você tiver downloads grandes ou scripts complexos, isso aumentará o tempo necessário para que a instância fique pronta para uso. Nesse caso, talvez seja necessário configurar um gancho do ciclo de vida para atrasar o alcance do estado `InService` de uma instância até que ela seja totalmente provisionada. Para obter mais informações sobre a adição de um gancho do ciclo de vida, ao seu grupo de Auto Scaling, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

Solicitar instâncias spot para aplicações flexíveis e com tolerância a falhas

Em seu modelo de execução, você tem a opção de solicitar instâncias spot sem data de encerramento ou duração. As instâncias spot do Amazon EC2 são capacidade de reserva disponível com grandes descontos em comparação com o preço do EC2 On-Demand. As Instâncias spot são uma opção econômica se houver flexibilidade quanto ao momento em que as aplicações serão executadas e se as aplicações poderão ser interrompidas. Para mais informações sobre como criar um modelo de execução que solicita instâncias spot, consulte Criar um modelo de execução usando configurações avançadas.

> **Important**
>
> As instâncias spot geralmente são usadas para complementar as instâncias sob demanda. Para este cenário, é possível especificar as mesmas configurações que são usadas no execução de instâncias spot como parte das configurações do grupo do Auto Scaling. Ao especificar as configurações como parte do grupo do Auto Scaling, você pode solicitar a execução de instâncias spot somente após a execução de um determinado número de instâncias sob demanda e, em seguida, continuar a iniciar alguma combinação de instâncias sob demanda e instâncias spot conforme o grupo for escalado. Para obter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

Este tópico descreve como iniciar apenas instâncias spot em seu grupo do Auto Scaling especificando configurações em um modelo de execução em vez de especificá-las no próprio grupo do Auto Scaling. As informações neste tópico também se aplicam a grupos do Auto Scaling que solicitem instâncias spot com uma configuração de execução. A diferença é que uma configuração de execução requer um preço máximo, mas para modelos de execução, o preço máximo é opcional.

Ao criar um ou modelo de execução para iniciar apenas instâncias spot, mantenha as seguintes considerações em mente:

- Preço spot. Você paga apenas o preço spot atual pelas instâncias spot que iniciar. Esse preço muda lentamente ao longo do tempo com base em tendências de oferta e demanda no longo prazo. Para mais informações, consulte Instâncias spot e Preços e economias no Guia do Usuário do Amazon EC2.
- Definir seu preço máximo. Você tem a opção de incluir um preço máximo por hora para instâncias spot no modelo de execução. Se seu preço máximo exceder o preço spot atual, o serviço do Amazon EC2 Spot atenderá à sua solicitação imediatamente mediante a disponibilidade de capacidade. Se o preço de instâncias spot ultrapassar o preço máximo para uma instância em execução em seu grupo do Auto Scaling, ele encerrará sua instância.

> **Warning**
>
> Talvez sua aplicação não seja executada se você não receber suas instâncias spot, como quando o preço máximo é muito baixo. Para aproveitar as instâncias spot disponíveis pelo maior tempo possível, defina seu preço máximo próximo ao preço sob demanda.

- Equilíbrio entre Zonas de disponibilidade. Se você especificar várias zonas de disponibilidade, o Amazon EC2 Auto Scaling distribuirá as solicitações spot entre as zonas especificadas. Se o preço máximo for muito baixo em uma zona de disponibilidade para que as solicitações sejam atendidas, o Amazon EC2 Auto Scaling verificará se elas foram atendidas nas outras zonas. Nesse caso, o Amazon EC2 Auto Scaling cancela as solicitações que falharam e as redistribui entre as zonas de disponibilidade com solicitações atendidas. Se o preço em uma zona de disponibilidade sem solicitações atendidas baixar o suficiente para que futuras solicitações tenham êxito, o Amazon EC2 Auto Scaling balanceará novamente entre todas as zonas de disponibilidade.
- Término de instância spot. As instâncias spot podem ser encerradas a qualquer momento. O serviço do Amazon EC2 Spot pode terminar instâncias spot em seu grupo do Auto Scaling conforme o preço ou a disponibilidade das instâncias spot mude. Ao escalar ou realizar verificação de integridade, o Amazon EC2 Auto Scaling também pode encerrar instâncias spot da mesma forma que pode terminar instâncias sob demanda. Quando uma instância é encerrada, qualquer armazenamento é excluído.
- Manter a capacidade desejada. Quando uma instância spot é encerrada, o Amazon EC2 Auto Scaling tenta iniciar outra instância spot para manter a capacidade desejada para o grupo. Se o preço spot atual for mais baixo que o preço máximo, uma instância spot será executada. Se a solicitação para uma instância spot não for bem-sucedida, ele continuará tentando.
- Alterar seu preço máximo. Para alterar o preço máximo, crie um novo modelo de execução ou atualize um modelo de execução existente com o novo preço máximo e, em seguida, associe-o a seu grupo do Auto Scaling. As instâncias spot existentes continuarão a ser executadas desde que o preço máximo especificado no modelo de execução usado para essas instâncias seja mais alto que o preço spot atual. Se você não definir um preço máximo, o preço máximo padrão será o preço sob demanda.

Use Capacity Blocks para workloads de machine learning.

Capacity Blocks ajuda você a reservar instâncias GPU altamente requisitadas em uma data futura para suportar suas workloads de machine learning (ML) de curta duração.

Para obter uma visão geral dos Capacity Blocks e saber como eles funcionam, consulte Capacity Blocks para ML no Guia do Usuário do Amazon EC2.

Para começar a usar Capacity Blocks, você cria uma reserva de capacidade em uma zona de disponibilidade específica. Os Capacity Blocks são entregues como reservas de capacidade `targeted` em uma única zona de disponibilidade. Ao criar o modelo de inicialização, especifique o ID da reserva e o tipo de instância do bloco de capacidade. Em seguida, atualize o grupo do Auto

Scaling para usar o modelo de inicialização que você criou e a zona de disponibilidade do bloco de capacidade. Quando a reserva do bloco de capacidade começar, use a escala programada para iniciar o mesmo número de instâncias que a reserva do bloco de capacidade.

> **Important**
>
> Capacity Blockssó estão disponíveis para determinados tipos de instância do Amazon EC2 e. Regiões da AWS Para mais informações, consulte Pré-requisitos no Guia do Usuário do Amazon EC2.

Conteúdo

- Diretrizes operacionais
- Especificar um bloco de capacidade em seu modelo de execução
- Limitações
- Recursos relacionados

Diretrizes operacionais

As diretrizes operacionais básicas a seguir devem ser seguidas por você ao usar um bloco de capacidade com um grupo do Auto Scaling.

- Reduza a escala horizontalmente do seu grupo do Auto Scaling até zero mais de 30 minutos antes do horário de término da reserva do bloco de capacidade. O Amazon EC2 encerrará todas as instâncias que ainda estiverem em execução 30 minutos antes do horário final do bloco de capacidade.
- Recomendamos que você use o dimensionamento agendado para aumentar a capacidade (adicionar instâncias) e diminuir a capacidade (remover instâncias) nos horários de reserva apropriados. Para obter mais informações, consulte Escalabilidade programada para o Amazon EC2 Auto Scaling.
- Adicione ganchos do ciclo de vida conforme necessário para realizar um desligamento suave da sua aplicação dentro das instâncias durante a redução da escala horizontalmente. Deixe tempo suficiente para que a ação do ciclo de vida seja concluída antes que o Amazon EC2 comece a encerrar forçosamente suas instâncias 30 minutos antes do horário de término da reserva do bloco de capacidade. Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

- Certifique-se de que o grupo do Auto Scaling aponte para a versão correta do modelo de execução durante toda a duração da reserva. Recomendamos apontar para uma versão específica do modelo de execução em vez da versão `$Default` ou `$Latest`.

> **Note**
>
> Se você deixar uma instância do bloco de capacidade em execução até o final da reserva e o Amazon EC2 a recuperar, as atividades de escala do seu grupo do Auto Scaling indicarão que ela estava "`taken out of service in response to an EC2 health check that indicated it had been terminated or stopped`", mesmo que tenha sido recuperada propositalmente no final do bloco de capacidade. Da mesma forma, o Amazon EC2 Auto Scaling tentará substituir a instância da mesma maneira que faz com qualquer instância que falhe em uma verificação de integridade. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Especificar um bloco de capacidade em seu modelo de execução

Para criar um modelo de execução que vise a um bloco de capacidade específico para seu grupo do Auto Scaling, use um dos métodos a seguir:

Console

Para especificar um bloco de execução no seu modelo de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. Na barra de navegação superior, selecione Região da AWS onde você criou seu bloco de capacidade.
3. No painel de navegação, escolha Instances e, em seguida, Launch Templates.
4. Escolha Criar modelo de execução e crie o modelo de execução. Inclua o ID da Imagem de máquina da Amazon (AMI), o tipo de instância e quaisquer outras configurações de execução conforme necessário.
5. Amplie a seção Detalhes avançados para ver as configurações avançadas.
6. Para a Opção de compra, escolha Blocos de capacidade.

7. Em Reserva de Capacidade, escolha Destino por ID e, em seguida, em Reserva de Capacidade - Destino por ID, escolha o ID de Reserva de Capacidade de um bloco de capacidade existente.
8. Quando terminar, selecione Criar modelo de execução.

   Para obter ajuda na criação de um grupo de Auto Scaling com um modelo de inicialização, consulte.Criar um grupo do Auto Scaling usando um modelo de execução

AWS CLI

Para especificar um bloco de capacidade em seu modelo de execução (AWS CLI)

Use o create-launch-templatecomando a seguir para criar um modelo de lançamento que especifica uma ID de reserva existente do Bloco de Capacidade. Substitua cada *user input placeholder* por suas próprias informações.

```
aws ec2 create-launch-template --launch-template-name my-template-for-capacity-block
  \
  --version-description AutoScalingVersion1 --region us-east-2 \
  --launch-template-data file://config.json
```

> **Tip**
>
> Se esse comando gerar um erro, verifique se você atualizou o AWS CLI localmente para a versão mais recente.

Conteúdo de `config.json`.

```
{
    "ImageId": "ami-04d5cc9b88example",
    "InstanceType": "p4d.24xlarge",
    "SecurityGroupIds": [
        "sg-903004f88example"
    ],
    "KeyName": "MyKeyPair",
    "InstanceMarketOptions": {
        "MarketType": "capacity-block"
    },
```

```
    "CapacityReservationSpecification": {
        "CapacityReservationTarget": {
            "CapacityReservationId": "cr-02168da1478b509e0"
        }
    }
}
```

O seguinte é um exemplo de saída.

```
{
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-068f72b724example",
        "LaunchTemplateName": "my-template-for-capacity-block",
        "CreateTime": "2023-10-27T15:12:44.000Z",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
    }
}
```

Você pode usar o describe-launch-template-versionscomando a seguir para verificar o ID de reserva do Capacity Block associado ao modelo de lançamento.

```
aws ec2 describe-launch-template-versions --launch-template-names my-template-for-capacity-block \
  --region us-east-2
```

A seguir está um exemplo de saída de um modelo de execução que especifica uma reserva de bloco de capacidade.

```
{
    "LaunchTemplateVersions": [
        {
            "LaunchTemplateId": "lt-068f72b724example",
            "LaunchTemplateName": "my-template-for-capacity-block",
            "VersionNumber": 1,
            "CreateTime": "2023-10-27T15:12:44.000Z",
            "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
            "DefaultVersion": true,
            "LaunchTemplateData": {
                "ImageId": "ami-04d5cc9b88example",
```

```
                "InstanceType": "p5.48xlarge",
                "SecurityGroupIds": [
                    "sg-903004f88example"
                ],
                "KeyName": "MyKeyPair",
                "InstanceMarketOptions": {
                    "MarketType": "capacity-block"
                },
                "CapacityReservationSpecification": {
                    "CapacityReservationTarget": {
                        "CapacityReservationId": "cr-02168da1478b509e0"
                    }
                }
            }
        }
    ]
}
```

Limitações

- O suporte para Capacity Blocks só está disponível se seu grupo do Auto Scaling tiver uma configuração compatível. Não há suporte para grupos de instâncias mistas e pools ativos.
- Você só pode selecionar um bloco de capacidade por vez.

Recursos relacionados

- Para obter os pré-requisitos e as recomendações para o uso de instâncias P5, consulte Introdução às instâncias P5 no Guia do usuário do Amazon EC2.
- O Amazon EKS suporta o uso Capacity Blocks para cargas de trabalho de machine learning (ML) de curta duração em clusters do Amazon EKS. Para obter mais informações, consulte para ML Capacity Blocks, no Guia do usuário do Amazon EKS.
- Você pode usar Capacity Blocks com tipos de instância e regiões compatíveis. No entanto, as reservas de capacidade sob demanda oferecem flexibilidade para reservar capacidade para outros tipos de instâncias e regiões. Para obter um tutorial que mostra como usar a opção Reserva de capacidade sob demanda, consulte Reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas.

Migre seus grupos do Auto Scaling para modelos de execução

Desde 1º de janeiro de 2023, não há mais suporte aos novos tipos de instância nas configurações de execução. Isso se aplica a qualquer tipo de instância adicionado Região da AWS após o lançamento inicial da região. Além disso, talvez você consiga criar uma configuração de execução com um tipo de instância que não tenha mais suporte em uma região. Para obter mais informações, consulte Configurações de execução do Auto Scaling.

Para migrar seus grupos de Auto Scaling de configurações de inicialização para modelos de inicialização, consulte as seguintes etapas.

> **Important**
>
> Certifique-se de que você tenha as permissões necessárias para trabalhar com modelos de execução. Para obter mais informações, consulte Permissões para trabalhar com modelos de inicialização.
> Certifique-se de que os recursos associados à configuração de execução, como grupos de segurança e perfis de instâncias do IAM, não sejam excluídos acidentalmente.
> Teste a operação de reversão para garantir que as alterações possam ser revertidas com segurança.

Etapa 1: encontrar grupos do Auto Scaling que usem configurações de execução

Para identificar se você tem grupos de Auto Scaling que ainda estão usando configurações de inicialização, execute o describe-auto-scaling-groupscomando a seguir usando o. AWS CLI*REGION*Substitua pelo seu Região da AWS.

```
aws autoscaling describe-auto-scaling-groups --region REGION \
  --query 'AutoScalingGroups[?LaunchConfigurationName!=`null`]'
```

O seguinte é um exemplo de saída.

```
[
    {
 "AutoScalingGroupName": "group-1",
 "AutoScalingGroupARN": "arn",
 "LaunchConfigurationName": "my-launch-config",
```

```
"MinSize": 1,
"MaxSize": 5,
"DesiredCapacity": 2,
"DefaultCooldown": 300,
"AvailabilityZones": [
          "us-west-2a",
          "us-west-2b",
          "us-west-2c"
       ],
"LoadBalancerNames": [],
"TargetGroupARNs": [],
"HealthCheckType": "EC2",
"HealthCheckGracePeriod": 300,
"Instances": [
          {
              "ProtectedFromScaleIn": false,
              "AvailabilityZone": "us-west-2a",
              "LaunchConfigurationName": "my-launch-config",
              "InstanceId": "i-05b4f7d5be44822a6",
              "InstanceType": "t3.micro",
              "HealthStatus": "Healthy",
              "LifecycleState": "InService"
          },
          {
              "ProtectedFromScaleIn": false,
              "AvailabilityZone": "us-west-2b",
              "LaunchConfigurationName": "my-launch-config",
              "InstanceId": "i-0c20ac468fa3049e8",
              "InstanceType": "t3.micro",
              "HealthStatus": "Healthy",
              "LifecycleState": "InService"
          }
],
"CreatedTime": "2023-03-09T22:15:11.611Z",
"SuspendedProcesses": [],
"VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
"EnabledMetrics": [],
"Tags": [
          {
 "ResourceId": "group-1",
 "ResourceType": "auto-scaling-group",
 "Key": "environment",
 "Value": "production",
 "PropagateAtLaunch": true
```

```
            }
        ],
 "TerminationPolicies": [
     "Default"
 ],
 "NewInstancesProtectedFromScaleIn": false,
 "ServiceLinkedRoleARN": "arn",
       "TrafficSources": []
    },

    ... additional groups ...

]
```

Como alternativa, para remover tudo, exceto os nomes dos grupos do Auto Scaling com os nomes de suas respectivas configurações de execução e tags na saída, execute o seguinte comando:

```
aws autoscaling describe-auto-scaling-groups --region REGION \
  --query 'AutoScalingGroups[?LaunchConfigurationName!=`null`].{AutoScalingGroupName:
 AutoScalingGroupName, LaunchConfigurationName: LaunchConfigurationName, Tags: Tags}'
```

Veja a seguir um exemplo de saída.

```
[
    {
        "AutoScalingGroupName": "group-1",
        "LaunchConfigurationName": "my-launch-config",
        "Tags": [
            {
                "ResourceId": "group-1",
                "ResourceType": "auto-scaling-group",
                "Key": "environment",
                "Value": "production",
                "PropagateAtLaunch": true
            }
        ]
    },

    ... additional groups ...

]
```

Para obter mais informações sobre filtragem, consulte Filtragem AWS CLI de saída no Guia do AWS Command Line Interface usuário.

Etapa 2: copiar uma configuração de execução para um modelo de execução

Você pode copiar uma configuração de execução para um modelo de execução usando o procedimento a seguir. Em seguida, você pode adicioná-lo ao seu grupo do Auto Scaling.

Copiar várias configurações de execução resulta em modelos de execução com nomes idênticos. Para alterar o nome dado a um modelo de execução durante o processo de cópia, é necessário copiar as configurações de execução uma a uma.

> **Note**
>
> O recurso de cópia só está disponível no console.

Para copiar uma configuração de execução para um modelo de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, em Ajuste de escala automático, escolha Grupos do Auto Scaling.
3. Escolha Configurações de execução na parte superior da página. Quando a confirmação for solicitada, escolha Exibir configurações de lançamento para confirmar que você deseja visualizar a página de Configurações de execução.
4. Selecione a configuração de execução que você deseja copiar e escolha Copy to launch template, Copy selected (Copiar para modelo de execução, Copiar selecionado). Um novo modelo de execução é criado com o mesmo nome e as mesmas opções da configuração de execução que você selecionou.
5. Em New launch template name (Novo nome de modelo de execução), você pode usar o nome da configuração de execução (o padrão) ou digitar um novo nome. Os nomes de modelo de execução devem ser exclusivos.
6. (Opcional) Selecione Criar um grupo do Auto Scaling usando o novo modelo.

   Você pode pular esta etapa para concluir a cópia da configuração de execução. Você não precisa criar um novo grupo do Auto Scaling.
7. Escolha Copiar.

Para copiar todas as configurações de execução para modelos de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, em Auto Scaling, escolha Launch Configurations (Configurações de execução).
3. Selecione Copy to launch template, Copy all (Copiar para modelo de execução, Copiar tudo). Isso copia cada configuração de execução na região atual para um novo modelo de execução com o mesmo nome e as mesmas opções.
4. Escolha Copiar.

Etapa 3: atualizar um grupo do Auto Scaling para usar um modelo de execução

Depois de criar um modelo de execução, você estará pronto para adicioná-lo ao seu grupo do Auto Scaling.

Para atualizar um grupo do Auto Scaling para usar um modelo de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página, mostrando informações sobre o grupo selecionado.
3. Na guia Details (Detalhes), escolha Launch configuration (Configuração de execução), Edit (Editar).
4. Escolha Switch to launch template (Alternar para modelo de execução).
5. Em Launch template (Modelo de execução), selecione seu modelo de execução.
6. Em Version (Versão), selecione a versão do modelo de execução, conforme necessário. Assim quer criar as versões do modelo de execução, poderá escolher se o grupo do Auto Scaling deve usar a versão padrão ou a versão mais recente do modelo de execução ao se ampliar.
7. Selecione Atualizar.

Para atualizar um grupo do Auto Scaling para usar um modelo de execução (AWS CLI)

O update-auto-scaling-groupcomando a seguir atualiza o grupo de Auto Scaling especificado para usar a versão inicial do modelo de execução especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-template-for-auto-scaling,Version='1'
```

Para obter exemplos de uso de comandos da CLI para atualizar um grupo do Auto Scaling para usar um modelo de execução, consulte Atualizar um grupo do Auto Scaling para usar um modelo de execução.

Etapa 4: substituir suas instâncias

Depois que você substituir a configuração de execução por um modelo de execução, todas as novas instâncias usarão o novo modelo de execução. As instâncias existentes não são afetadas.

Para atualizar as instâncias existentes, você pode iniciar uma atualização de instância para substituir as instâncias em seu grupo do Auto Scaling em vez de substituir manualmente algumas instâncias de cada vez. Para obter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling. Se o grupo for grande, uma atualização de instância pode ser particularmente útil.

Como alternativa, você pode permitir que a escalabilidade automática substitua gradualmente as instâncias existentes por novas instâncias com base nas políticas de encerramento do grupo, ou você pode encerrá-las. O encerramento manual força seu grupo do Auto Scaling a lançar novas instâncias para manter a capacidade desejada do grupo. Para obter mais informações, consulte Terminar uma instância no Guia do usuário do Amazon EC2.

Mais informações

Para obter mais informações, consulte O Amazon EC2 Auto Scaling não adicionará mais suporte para novos recursos do EC2 às configurações AWS de lançamento no blog de computação.

Para ver um tópico que mostra como migrar AWS CloudFormation pilhas de configurações de lançamento para modelos de execução, consulte. Migre AWS CloudFormation pilhas para modelos de lançamento

Migre AWS CloudFormation pilhas para modelos de lançamento

Você pode migrar seus modelos de CloudFormation pilha existentes das configurações de lançamento para os modelos de lançamento. Para fazer isso, adicione um modelo de execução

diretamente a um modelo de pilha existente e, em seguida, associe o modelo de execução ao grupo do Auto Scaling no modelo de pilha. Em seguida, use seu modelo modificado para atualizar sua pilha.

Ao migrar para modelos de lançamento, este tópico economiza seu tempo ao fornecer instruções para reescrever as configurações de lançamento em seus modelos de CloudFormation pilha como modelos de lançamento. Para obter mais informações sobre migração de configurações de lançamento para modelos de execução, consulte Migre seus grupos do Auto Scaling para modelos de execução.

Tópicos

- Encontre grupos do Auto Scaling que usam uma configuração de execução
- Atualizar uma pilha para usar um modelo de execução
- Compreender atualização de comportamentos de recursos da pilha
- Rastrear a migração
- Referência do mapeamento de configuração de execução

Encontre grupos do Auto Scaling que usam uma configuração de execução

Para localizar grupos do Auto Scaling que usam uma configuração de execução

- Use o describe-auto-scaling-groupscomando a seguir para listar os nomes dos grupos do Auto Scaling que estão usando configurações de inicialização na região especificada. Inclua a `--filters` opção de restringir os resultados aos grupos associados a uma CloudFormation pilha (filtrando pela chave da `aws:cloudformation:stack-name` tag).

```
aws autoscaling describe-auto-scaling-groups --region REGION \
  --filters Name=tag-key,Values=aws:cloudformation:stack-name \
  --query 'AutoScalingGroups[?LaunchConfigurationName!
=`null`].AutoScalingGroupName'
```

  Veja a seguir um exemplo de saída.

```
[
    "{stack-name}-group-1",
    "{stack-name}-group-2",
    "{stack-name}-group-3"
```

```
]
```

Você pode encontrar outros AWS CLI comandos úteis para encontrar grupos do Auto Scaling para migrar e filtrar a saída. Migre seus grupos do Auto Scaling para modelos de execução

> **Important**
>
> Se os recursos da sua pilha tiverem `AWSEB` em seu nome, isso significa que eles foram criados por meio AWS Elastic Beanstalk de. Nesse caso, é necessário atualizar o ambiente do Beanstalk para que o Elastic Beanstalk remova a configuração de execução e a substitua por um modelo de execução.

Atualizar uma pilha para usar um modelo de execução

Siga as etapas desta seção para fazer o seguinte:

- Reescreva a configuração de execução como um modelo de execução usando as propriedades equivalentes do modelo de execução.
- Associe o novo modelo de execução ao grupo do Auto Scaling.
- Implemente essas atualizações.

Para modificar o modelo da pilha e atualizar a pilha

1. Siga os mesmos procedimentos gerais para modificar o modelo de pilha descritos em Modificar um modelo de pilha no Guia do usuário AWS CloudFormation .
2. Reescreva a configuração de execução como um modelo de execução. Veja o exemplo a seguir:

   Exemplo: uma configuração de execução simples

   ```
   ---
   Resources:
     myLaunchConfig:
       Type: AWS::AutoScaling::LaunchConfiguration
       Properties:
         ImageId: ami-02354e95b3example
         InstanceType: t3.micro
   ```

```
      SecurityGroups:
        - !Ref EC2SecurityGroup
      KeyName: MyKeyPair
      BlockDeviceMappings:
        - DeviceName: /dev/xvda
          Ebs:
            VolumeSize: 150
            DeleteOnTermination: true
      UserData:
        Fn::Base64: !Sub |
          #!/bin/bash -xe
          yum install -y aws-cfn-bootstrap
          /opt/aws/bin/cfn-signal -e $? --stack ${AWS::StackName} --resource myASG
 --region ${AWS::Region}
```

Exemplo: O modelo de execução equivalente

```
---
Resources:
  myLaunchTemplate:
    Type: AWS::EC2::LaunchTemplate
    Properties:
      LaunchTemplateName: !Sub ${AWS::StackName}-launch-template
      LaunchTemplateData:
        ImageId: ami-02354e95b3example
        InstanceType: t3.micro
        SecurityGroupIds:
          - Ref! EC2SecurityGroup
        KeyName: MyKeyPair
        BlockDeviceMappings:
          - DeviceName: /dev/xvda
            Ebs:
              VolumeSize: 150
              DeleteOnTermination: true
        UserData:
          Fn::Base64: !Sub |
            #!/bin/bash -x
            yum install -y aws-cfn-bootstrap
            /opt/aws/bin/cfn-signal -e $? --stack ${AWS::StackName} --resource
 myASG --region ${AWS::Region}
```

Para obter informações de referência sobre todas as propriedades que o Amazon EC2 suporta, consulte o AWS::EC2::LaunchTemplateGuia do AWS CloudFormation usuário.

Observe como o modelo de execução inclui a propriedade `LaunchTemplateName` com um valor de `!Sub ${AWS::StackName}-launch-template`. Isso é necessário se você quiser que o nome do modelo de execução inclua o nome da pilha.

3. Se a propriedade **`IamInstanceProfile`** estiver presente em sua configuração de execução, será necessário convertê-la em uma estrutura e especificar o nome ou o ARN do perfil de instância. Para ver um exemplo, consulte AWS::EC2::LaunchTemplate.
4. Se as propriedades,**`AssociatePublicIpAddress`** **`InstanceMonitoring`** ou **`PlacementTenancy`** estiverem presentes em sua configuração de execução, será necessário convertê-las em uma estrutura. Para obter exemplos, consulte AWS::EC2::LaunchTemplate.

   Uma exceção é quando o valor da propriedade `MapPublicIpOnLaunch` nas sub-redes que você usou para seu grupo do Auto Scaling coincide com o valor da propriedade `AssociatePublicIpAddress` em sua configuração de execução. Neste caso, você pode ignorar a propriedade`AssociatePublicIpAddress`. A `AssociatePublicIpAddress` propriedade só é usada para substituir a `MapPublicIpOnLaunch` propriedade e alterar se as instâncias recebem um IPv4 endereço público na inicialização.
5. Você pode copiar grupos de segurança da propriedade **`SecurityGroups`** para um dos dois lugares em seu modelo de execução. Normalmente, você copia os grupos de segurança para a propriedade`SecurityGroupIds`. No entanto, se você criar uma estrutura `NetworkInterfaces` em seu modelo de execução para especificar a propriedade,`AssociatePublicIpAddress` será necessário copiar os grupos de segurança para a propriedade `Groups` da interface de rede.
6. Se alguma estrutura `BlockDeviceMapping` estiver presente em sua configuração de execução com **`NoDevice`** definido como,`true` será necessário especificar uma string vazia para `NoDevice` em seu modelo de execução para que o Amazon EC2 omita o dispositivo.
7. Se a propriedade **`SpotPrice`** estiver presente em sua configuração de execução, recomendamos que você a omita do seu modelo de execução. Suas instâncias spot serão executadas pelo preço spot atual. Este preço nunca excederá o preço sob demanda.

   Para solicitar instâncias spot, você tem duas opções mutuamente exclusivas:

- A primeira é usar a estrutura `InstanceMarketOptions` em seu modelo de execução (não recomendado). Para obter mais informações, consulte AWS::EC2::LaunchTemplate InstanceMarketOptionso Guia AWS CloudFormation do usuário.
- A outra é adicionar uma estrutura `MixedInstancesPolicy` ao seu grupo do Auto Scaling. Isto oferece mais opções de como você faz a solicitação. Uma solicitação de instância spot em seu modelo de execução não é compatível com mais de uma seleção de tipo de instância por grupo do Auto Scaling. No entanto, uma política de instâncias mistas oferece suporte a mais de uma seleção de tipo de instância por grupo do Auto Scaling. As solicitações de Instância spot se beneficiam de ter mais de um tipo de instância para escolher. Para obter mais informações, consulte AWS::AutoScaling::AutoScaling MixedInstancesPolicyAWS::AutoScaling::AutoScalingGrupo no Guia AWS CloudFormation do usuário.

8. Remova a **`LaunchConfigurationName`** propriedade do recurso AWS::AutoScaling::AutoScalingAWS::AutoScaling::AutoScalingGrupo . Adicione o modelo de execução em seu lugar.

   Nos exemplos a seguir, a função intrínseca Ref obtém o ID do AWS::EC2::LaunchTemplaterecurso com o ID lógico. `myLaunchTemplate` A GetAttfunção obtém o número da versão mais recente (por exemplo,1) do modelo de lançamento da `Version` propriedade.

   Exemplo: sem uma política de instâncias mistas

   ```
   ---
   Resources:
     myASG:
       Type: AWS::AutoScaling::AutoScalingGroup
       Properties:
         LaunchTemplate:
           LaunchTemplateId: !Ref myLaunchTemplate
           Version: !GetAtt myLaunchTemplate.LatestVersionNumber
         ...
   ```

   Exemplo: com uma política de instâncias mistas

   ```
   ---
   Resources:
     myASG:
   ```

```
    Type: AWS::AutoScaling::AutoScalingGroup
    Properties:
      MixedInstancesPolicy:
        LaunchTemplate:
          LaunchTemplateSpecification:
            LaunchTemplateId: !Ref myLaunchTemplate
            Version: !GetAtt myLaunchTemplate.LatestVersionNumber
      ...
```

Para obter informações de referência sobre todas as propriedades que o Amazon EC2 Auto Scaling suporta AWS::AutoScaling::AutoScaling, AWS::AutoScaling::AutoScaling Group Group AWS CloudFormation no Guia do usuário.

9. Quando você estiver pronto para implantar essas atualizações, siga os CloudFormation procedimentos para atualizar a pilha com seu modelo de pilha modificado. Para obter mais informações, consulte Modificar um modelo de pilha no Guia do usuário AWS CloudFormation .

Compreender atualização de comportamentos de recursos da pilha

CloudFormation atualiza os recursos da pilha comparando as alterações entre o modelo atualizado que você fornece e as configurações de recursos que você descreveu na versão anterior do seu modelo de pilha. As configurações de recursos que não foram alteradas permanecem inalteradas durante o processo de atualização.

CloudFormation suporta o UpdatePolicyatributo para grupos de Auto Scaling. Durante uma atualização, se `UpdatePolicy` estiver definido como`AutoScalingRollingUpdate`, CloudFormation substitui `InService` as instâncias após você executar as etapas desse procedimento. Se `UpdatePolicy` estiver definido como`AutoScalingReplacingUpdate`, CloudFormation substitui o grupo Auto Scaling e sua piscina aquecida (se houver).

Se você não especificou um `UpdatePolicy` atributo para seu grupo de Auto Scaling, o modelo de lançamento é verificado quanto à exatidão, mas CloudFormation não implanta nenhuma alteração nas instâncias do grupo Auto Scaling. Todas as novas instâncias receberão seu modelo de execução, mas as instâncias existentes continuarão a ser executadas com a configuração de execução com a qual foram executadas originalmente (apesar da inexistência da configuração de execução). A exceção é quando você altera suas opções de compra, por exemplo, adicionando uma política de instâncias mistas. Neste caso, seu grupo do Auto Scaling substitui gradualmente as instâncias existentes por novas instâncias para corresponder às novas opções de compra.

Se você precisar reverter uma alteração para passar de configurações de execução para modelos de execução, certifique-se de testar a operação de reversão.

Rastrear a migração

Para rastrear a migração

1. No console do CloudFormation, selecione a pilha que você atualizou e, em seguida, escolha a guia Eventos para visualizar eventos de pilhas.
2. Para atualizar a lista de eventos com os eventos mais recentes, escolha o botão Atualizar no CloudFormation console.
3. Enquanto sua pilha estiver sendo atualizada, você notará vários eventos para cada atualização de recurso. Se você ver uma exceção na coluna Motivo do status que indica um problema ao tentar criar o modelo de execução, consulte Solucionar problemas do Amazon EC2 Auto Scaling: modelos de execução para as possíveis causas.
4. (Opcional) Dependendo do uso do atributo,`UpdatePolicy` você pode monitorar o progresso do seu grupo do Auto Scaling na página de grupos do Auto Scaling do console do Amazon EC2. Selecione o grupo do Auto Scaling. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status (Status) mostra se o seu grupo do Auto Scaling iniciou ou encerrou instâncias com êxito, ou se a ação de escalabilidade ainda está em andamento.
5. Quando a atualização da pilha estiver concluída, CloudFormation emite um evento de `UPDATE_COMPLETE` pilha. Para obter mais informações, consulte Monitorar o andamento de uma atualização de pilha no Guia do usuário AWS CloudFormation .
6. Depois que a atualização da pilha estiver concluída, abra a página de modelos de execução e a página de configurações de execução do console do Amazon EC2. Você notará que um novo modelo de execução foi criado e a configuração de execução foi excluída.

Referência do mapeamento de configuração de execução

Para fins de referência, a tabela a seguir lista todas as propriedades de nível superior no AWS::AutoScaling::LaunchConfigurationrecurso com suas propriedades correspondentes no AWS::EC2::LaunchTemplaterecurso.

Propriedade da fonte de configuração de execução	Propriedade target do modelo de execução
`AssociatePublicIpAddress`	`NetworkInterfaces.AssociatePublicIpAddress`
`BlockDeviceMappings`	`BlockDeviceMappings`
`ClassicLinkVPCId`	Não disponível[1]
`ClassicLinkVPCSecurityGroups`	Não disponível[1]
`EbsOptimized`	`EbsOptimized`
`IamInstanceProfile`	Especifique `IamInstanceProfile.Arn` ou `IamInstanceProfile.Name`, mas não ambos
`ImageId`	`ImageId`
`InstanceId`	`InstanceId`
`InstanceMonitoring`	`Monitoring.Enabled`
`InstanceType`	`InstanceType`
`KernelId`	`KernelId`
`KeyName`	`KeyName`
`LaunchConfigurationName`	`LaunchTemplateName`
`MetadataOptions`	`MetadataOptions`
`PlacementTenancy`	`Placement.Tenancy`
`RamDiskId`	`RamDiskId`
`SecurityGroups`	Especifique `SecurityGroupIds` ou `NetworkInterfaces.Groups`, mas não ambos

Propriedade da fonte de configuração de execução	Propriedade target do modelo de execução
`SpotPrice`	`InstanceMarketOptions.SpotOptions.MaxPrice`
`UserData`	`UserData`

[1] As propriedades `ClassicLinkVPCId` e `ClassicLinkVPCSecurityGroups` não estão disponíveis para uso em um modelo de execução porque o EC2-Classic não está mais disponível.

Exemplos para criar e gerenciar modelos de lançamento com o AWS CLI

Você pode criar e gerenciar modelos de lançamento por meio do Console de gerenciamento da AWS, AWS Command Line Interface (AWS CLI) ou SDKs. Esta seção mostra exemplos de criação e gerenciamento de modelos de lançamento para o Amazon EC2 Auto AWS CLI Scaling a partir do.

Conteúdo

- Exemplo de uso
- Criar um modelo de execução básico
- Especificar etiquetas que marcam instâncias ao iniciar
- Especificar um perfil do IAM a ser transmitido às instâncias
- Atribuir um endereço IP público
- Especificar um script de dados do usuário que configura instâncias ao iniciar
- Especificar um mapeamento de dispositivos de blocos
- Especificar hosts dedicados para trazer licenças de software de fornecedores externos
- Especificar uma interface de rede existente
- Criar várias interfaces de rede
- Gerenciar modelos de execução
- Atualizar um grupo do Auto Scaling para usar um modelo de execução

Exemplo de uso

```
{
    "LaunchTemplateName": "my-template-for-auto-scaling",
    "VersionDescription": "test description",
    "LaunchTemplateData": {
        "ImageId": "ami-04d5cc9b88example",
        "InstanceType": "t2.micro",
        "SecurityGroupIds": [
            "sg-903004f88example"
        ],
        "KeyName": "MyKeyPair",
        "Monitoring": {
            "Enabled": true
        },
        "Placement": {
            "Tenancy": "dedicated"
        },
        "CreditSpecification": {
            "CpuCredits": "unlimited"
        },
        "MetadataOptions": {
            "HttpTokens": "required",
            "HttpPutResponseHopLimit": 1,
            "HttpEndpoint": "enabled"
        }
    }
}
```

Criar um modelo de execução básico

Para criar um modelo de lançamento básico, use o create-launch-templatecomando da seguinte forma, com estas modificações:

- Substitua `ami-04d5cc9b88example` pelo ID da AMI a partir da qual as instâncias serão inicializadas.
- Substitua `t2.micro` por um tipo de instância compatível com a AMI especificada.

Este exemplo cria um modelo de lançamento com o nome*`my-template-for-auto-scaling`*. Se as instâncias criadas por esse modelo de execução forem executadas em uma VPC padrão, elas

receberão um endereço IP público por padrão. Se as instâncias forem executadas em uma VPC não padrão, elas não receberão um endereço IP público por padrão.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data
 '{"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.micro"}'
```

Para obter mais informações sobre como citar parâmetros formatados em JSON, consulte Uso de aspas com strings na AWS CLI no Manual do usuário da AWS Command Line Interface .

Como alternativa, é possível especificar os parâmetros formatados em JSON em um arquivo de configuração.

O exemplo a seguir cria um modelo de execução básico, fazendo referência a um arquivo de configuração para valores de parâmetro de modelo de execução.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data file://config.json
```

Conteúdo de `config.json`:

```
{
    "ImageId":"ami-04d5cc9b88example",
    "InstanceType":"t2.micro"
}
```

Especificar etiquetas que marcam instâncias ao iniciar

O exemplo a seguir adiciona uma tag (por exemplo, `purpose=webserver`) a instâncias na execução.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"TagSpecifications":[{"ResourceType":"instance","Tags":
[{"Key":"purpose","Value":"webserver"}]}],"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.
```

> **Note**
>
> Se você especificar tags de instância em seu modelo de execução e optar por propagar tags de seu grupo do Auto Scaling para suas instâncias, todas as tags serão mescladas. Se a mesma chave da etiqueta for especificada para uma etiqueta no modelo de execução e uma etiqueta no grupo do Auto Scaling, então, o valor da etiqueta do grupo terá precedência.

Especificar um perfil do IAM a ser transmitido às instâncias

O exemplo a seguir especifica o nome do perfil da instância associada ao perfil do IAM a ser passado às instâncias na execução. Para obter mais informações, consulte Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
--launch-template-data '{"IamInstanceProfile":{"Name":"my-instance-
profile"},"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.micro"}'
```

Atribuir um endereço IP público

O create-launch-templateexemplo a seguir configura o modelo de execução para atribuir endereços públicos às instâncias executadas em uma VPC não padrão.

> **Note**
>
> Quando você especificar uma interface de rede, especifique um valor para `Groups` que corresponda aos grupos de segurança da VPC nos quais seu grupo do Auto Scaling iniciará instâncias. Especifique as sub-redes da VPC como propriedades do grupo do Auto Scaling.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"NetworkInterfaces":
[{"DeviceIndex":0,"AssociatePublicIpAddress":true,"Groups":
["sg-903004f88example"],"DeleteOnTermination":true}],"ImageId":"ami-04d5cc9b88example","Instanc
```

Especificar um script de dados do usuário que configura instâncias ao iniciar

O exemplo a seguir especifica um script de dados do usuário como uma string codificada em base64 que configura instâncias na execução. O create-launch-templatecomando requer dados do usuário codificados em base64.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
--launch-template-data
 '{"UserData":"IyEvYmluL2Jhc...","ImageId":"ami-04d5cc9b88example","InstanceType":"t2.micro"}'
```

Especificar um mapeamento de dispositivos de blocos

O create-launch-templateexemplo a seguir cria um modelo de execução com um mapeamento de dispositivos de blocos: um volume do EBS de 22 gigabytes mapeado para. `/dev/xvdcz` O volume `/dev/xvdcz` usa o tipo de volume SSD de uso geral (gp2) e é excluído ao terminar a instância à qual ele está anexado.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"BlockDeviceMappings":[{"DeviceName":"/dev/xvdcz","Ebs":
{"VolumeSize":22,"VolumeType":"gp2","DeleteOnTermination":true}}],"ImageId":"ami-04d5cc9b88exam
```

Especificar hosts dedicados para trazer licenças de software de fornecedores externos

Se você especificar locação de host, você pode especificar um grupo de recursos de host e uma configuração licenças de License Manager para trazer licenças de software qualificáveis de fornecedores externos. Em seguida, você pode usar as licenças em instâncias do EC2 usando o comando a seguir create-launch-template.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"Placement":
{"Tenancy":"host","HostResourceGroupArn":"arn"},"LicenseSpecifications":
[{"LicenseConfigurationArn":"arn"}],"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.micro"
```

Especificar uma interface de rede existente

O create-launch-templateexemplo a seguir configura a interface de rede primária para usar uma interface de rede existente.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"NetworkInterfaces":
[{"DeviceIndex":0,"NetworkInterfaceId":"eni-
b9a5ac93","DeleteOnTermination":false}],"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.mi
```

Criar várias interfaces de rede

O create-launch-templateexemplo a seguir adiciona uma interface de rede secundária. A interface de rede primária tem um índice de dispositivo de 0, e a interface de rede secundária tem um índice de dispositivo de 1.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"NetworkInterfaces":[{"DeviceIndex":0,"Groups":
["sg-903004f88example"],"DeleteOnTermination":true},{"DeviceIndex":1,"Groups":
["sg-903004f88example"],"DeleteOnTermination":true}],"ImageId":"ami-04d5cc9b88example","Instanc
```

Se você usar um tipo de instância compatível com várias placas de rede e adaptadores Elastic Fabric (EFAs), poderá adicionar uma interface secundária a uma placa de rede secundária e habilitar o EFA usando o comando a seguir create-launch-template. Para obter mais informações, consulte Adição de um EFA a um modelo de execução no Guia do usuário do Amazon EC2.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling --
version-description version1 \
  --launch-template-data '{"NetworkInterfaces":
[{"NetworkCardIndex":0,"DeviceIndex":0,"Groups":
["sg-7c2270198example"],"InterfaceType":"efa","DeleteOnTermination":true},
{"NetworkCardIndex":1,"DeviceIndex":1,"Groups":
["sg-7c2270198example"],"InterfaceType":"efa","DeleteOnTermination":true}],"ImageId":"ami-09d95
```

> **Warning**
>
> O tipo de instância p4d.24xlarge incorre em custos mais altos do que os outros exemplos desta seção. Para obter mais informações sobre Ipreços de instâncias P4d, consulte Preços de instâncias P4d do Amazon EC2.

> **Note**
>
> Anexar várias interfaces de rede da mesma sub-rede a uma instância pode introduzir roteamento assimétrico, especialmente em instâncias que us~em uma variante do Linux que não seja da Amazon. Se você precisar desse tipo de configuração, deverá configurar a interface de rede secundária dentro do sistema operacional. Para ver um exemplo, consulte Como posso fazer minha interface de rede secundária funcionar na minha instância do Ubuntu EC2? no Centro de AWS Conhecimento.

Gerenciar modelos de execução

AWS CLI Isso inclui vários outros comandos que ajudam você a gerenciar seus modelos de lançamento.

Conteúdo

- Listar e descrever modelos de execução
- Criar uma versão de modelo de execução
- Excluir uma versão de modelo de execução
- Excluir um modelo de execução

Listar e descrever modelos de execução

Você pode usar dois AWS CLI comandos para obter informações sobre seus modelos de lançamento: describe-launch-templatesdescribe-launch-template-versionse.

O describe-launch-templatescomando permite que você obtenha uma lista de qualquer um dos modelos de lançamento que você criou. Você pode usar uma opção para filtrar resultados em um nome de modelo de execução, tempo de criação, chave de tag ou combinação de chave-valor de

tag. Esse comando retorna informações resumidas sobre qualquer um dos modelos de execução, incluindo o identificador de modelo de execução, a versão mais recente e a versão padrão.

O exemplo a seguir fornece um resumo do modelo de execução especificado.

```
aws ec2 describe-launch-templates --launch-template-names my-template-for-auto-scaling
```

O seguinte é um exemplo de resposta.

```
{
    "LaunchTemplates": [
        {
            "LaunchTemplateId": "lt-068f72b729example",
            "LaunchTemplateName": "my-template-for-auto-scaling",
            "CreateTime": "2020-02-28T19:52:27.000Z",
            "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
            "DefaultVersionNumber": 1,
            "LatestVersionNumber": 1
        }
    ]
}
```

Se você não usar a opção `--launch-template-names` para limitar a saída a um modelo de execução, informações sobre todos os modelos de execução serão retornadas.

O describe-launch-template-versionscomando a seguir fornece informações que descrevem as versões do modelo de lançamento especificado.

```
aws ec2 describe-launch-template-versions --launch-template-id lt-068f72b729example
```

O seguinte é um exemplo de resposta.

```
{
    "LaunchTemplateVersions": [
        {
            "VersionDescription": "version1",
            "LaunchTemplateId": "lt-068f72b729example",
            "LaunchTemplateName": "my-template-for-auto-scaling",
            "VersionNumber": 1,
            "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
```

```
            "LaunchTemplateData": {
                "TagSpecifications": [
                    {
                        "ResourceType": "instance",
                        "Tags": [
                            {
                                "Key": "purpose",
                                "Value": "webserver"
                            }
                        ]
                    }
                ],
                "ImageId": "ami-04d5cc9b88example",
                "InstanceType": "t2.micro",
                "NetworkInterfaces": [
                    {
                        "DeviceIndex": 0,
                        "DeleteOnTermination": true,
                        "Groups": [
                            "sg-903004f88example"
                        ],
                        "AssociatePublicIpAddress": true
                    }
                ]
            },
            "DefaultVersion": true,
            "CreateTime": "2020-02-28T19:52:27.000Z"
        }
    ]
}
```

Criar uma versão de modelo de execução

O create-launch-template-versioncomando a seguir cria uma nova versão do modelo de lançamento com base na versão 1 do modelo de execução e especifica uma ID de AMI diferente.

```
aws ec2 create-launch-template-version --launch-template-id lt-068f72b729example --
version-description version2 \
  --source-version 1 --launch-template-data "ImageId=ami-c998b6b2example"
```

Para definir a versão padrão do modelo de lançamento, use o modify-launch-templatecomando.

Excluir uma versão de modelo de execução

O delete-launch-template-versionscomando a seguir exclui a versão do modelo de lançamento especificado.

```
aws ec2 delete-launch-template-versions --launch-template-id lt-068f72b729example --
versions 1
```

Excluir um modelo de execução

Se você não precisar mais de um modelo de lançamento, poderá excluí-lo usando o delete-launch-templatecomando a seguir. A exclusão de um modelo de execução excluirá todas as suas versões.

```
aws ec2 delete-launch-template --launch-template-id lt-068f72b729example
```

Atualizar um grupo do Auto Scaling para usar um modelo de execução

Você pode usar o update-auto-scaling-groupcomando para adicionar um modelo de lançamento a um grupo de Auto Scaling existente.

Atualizar um grupo do Auto Scaling para usar a versão mais recente de um modelo de execução

O update-auto-scaling-groupcomando a seguir atualiza o grupo de Auto Scaling especificado para usar a versão mais recente do modelo de execução especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateId=lt-068f72b729example,Version='$Latest'
```

Atualizar um grupo do Auto Scaling para usar uma versão específica de um modelo de execução

O update-auto-scaling-groupcomando a seguir atualiza o grupo de Auto Scaling especificado para usar uma versão específica do modelo de execução especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-template-for-auto-scaling,Version='2'
```

Use AWS Systems Manager parâmetros em vez de AMI IDs em modelos de lançamento

Esta seção mostra como criar um modelo de execução que especifica um parâmetro da AWS Systems Manager que faz referência a um ID de imagem de máquina da Amazon (AMI). Você pode usar um parâmetro armazenado no mesmo Conta da AWS, um parâmetro compartilhado de outro Conta da AWS ou um parâmetro público para uma AMI pública mantida pela AWS.

Com os parâmetros do Systems Manager, você pode atualizar seus grupos de Auto Scaling para usar a nova AMI IDs sem precisar criar novos modelos de execução ou novas versões dos modelos de execução sempre que uma ID de AMI for alterada. Eles IDs podem mudar regularmente, como quando uma AMI é atualizada com as atualizações mais recentes do sistema operacional ou do software.

Você pode criar, atualizar ou excluir seus próprios parâmetros do Systems Manager usando o Parameter Store, um recurso de AWS Systems Manager. É necessário criar um parâmetro do Systems Manager para usá-lo em um modelo de execução. Para começar, você pode criar um parâmetro com o tipo de dados `aws:ec2:image` e, no valor, inserir o ID de uma AMI. O ID de AMI tem o formato `ami-<identifier>`, por exemplo, `ami-123example456`. O ID de AMI correto depende do tipo de instância e da Região da AWS na qual você está iniciando o grupo do Auto Scaling.

Para obter mais informações sobre como criar um parâmetro válido para um ID de AMI, consulte Crie um parâmetro do Systems Manager.

Criar um modelo de execução que especifique um parâmetro para a AMI

Para criar um modelo de execução que especifique um parâmetro para a AMI, use um dos métodos a seguir:

Console

Para criar um modelo de lançamento usando um AWS Systems Manager parâmetro

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, escolha Launch Templates (Modelos de execução) e Create launch template (Criar modelo de execução).
3. Em Device template name (Nome do modelo de dispositivo), insira um nome descritivo para o modelo.

4. Em Imagens do aplicativo e do sistema operacional (Amazon Machine Image), escolha Procurar mais AMIs.
5. Escolha o botão de seta à direita da barra de pesquisa e escolha Especificar parâmetro personalizado value/Systems do Gerenciador.
6. Na caixa de diálogo Especificar valor personalizado ou parâmetro do Systems Manager, faça o seguinte:

   a. Em ID de AMI ou string de parâmetros do Systems Manager, insira o nome do parâmetro do Systems Manager usando um destes formatos:

      Para fazer referência a um parâmetro público:

      - **resolve:ssm:*public-parameter***

      Para fazer referência a um parâmetro armazenado na mesma conta:

      - **resolve:ssm:*parameter-name***
      - **resolve:ssm:*parameter-name*:*version-number***
      - **resolve:ssm:*parameter-name*:*label***

      Para fazer referência a um parâmetro compartilhado de outra Conta da AWS:

      - **resolve:ssm:*parameter-ARN***
      - **resolve:ssm:*parameter-ARN*:*version-number***
      - **resolve:ssm:*parameter-ARN*:*label***

   b. Escolha Salvar.
7. Configure qualquer outro parâmetro do modelo de execução, conforme necessário, e escolha Criar modelo de execução. Para obter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.

AWS CLI

Para criar um modelo de execução que especifique um parâmetro do Systems Manager, você pode usar um dos seguintes comandos de exemplo. Substitua cada *`user input placeholder`* por suas próprias informações.

Exemplo: criar um modelo de lançamento que especifique um parâmetro público AWS de propriedade

Use a seguinte sintaxe: `resolve:ssm:public-parameter`, em que `resolve:ssm` é o prefixo padrão e *`public-parameter`* é o caminho e o nome do parâmetro público.

Neste exemplo, o modelo de execução usa um parâmetro público AWS fornecido para iniciar instâncias usando a AMI mais recente do Amazon Linux 2 Região da AWS que está configurada para seu perfil.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling
 --version-description version1 \
  --launch-template-data file://config.json
```

Conteúdo de `config.json`:

```
{
    "ImageId":"resolve:ssm:/aws/service/ami-amazon-linux-latest/amzn2-ami-hvm-
x86_64-gp2",
    "InstanceType":"t2.micro"
}
```

O seguinte é um exemplo de resposta.

```
{
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-089c023a30example",
        "LaunchTemplateName": "my-template-for-auto-scaling",
        "CreateTime": "2022-12-28T19:52:27.000Z",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
    }
}
```

Exemplo: Criar um modelo de execução que especifique um parâmetro armazenado na mesma conta

Use a seguinte sintaxe: `resolve:ssm:parameter-name`, em que `resolve:ssm` é o prefixo padrão e *`parameter-name`* é o nome do parâmetro do Systems Manager.

O exemplo a seguir cria um modelo de execução que obtém o ID de AMI de um parâmetro do Systems Manager existente chamado *golden-ami*.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling \
  --launch-template-data file://config.json
```

Conteúdo de `config.json`:

```
{
    "ImageId":"resolve:ssm:golden-ami",
    "InstanceType":"t2.micro"
}
```

Quando nenhuma versão é especificada, a versão padrão do parâmetro é a versão mais recente.

O exemplo a seguir referencia uma versão específica do parâmetro *golden-ami*. O exemplo usa a versão *3* do parâmetro *golden-ami*, mas é possível usar qualquer número de versão válido.

```
{
    "ImageId":"resolve:ssm:golden-ami:3",
    "InstanceType":"t2.micro"
}
```

O exemplo semelhante a seguir referencia o rótulo de parâmetro *prod* que é mapeado em uma versão específica do parâmetro *golden-ami*.

```
{
    "ImageId":"resolve:ssm:golden-ami:prod",
    "InstanceType":"t2.micro"
}
```

O seguinte é um exemplo de saída.

```
{
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-068f72b724example",
        "LaunchTemplateName": "my-template-for-auto-scaling",
        "CreateTime": "2022-12-27T17:11:21.000Z",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "DefaultVersionNumber": 1,
```

```
        "LatestVersionNumber": 1
    }
}
```

Exemplo: criar um modelo de lançamento que especifique um parâmetro compartilhado de outro Conta da AWS

Use a seguinte sintaxe: `resolve:ssm:`*`parameter-ARN`*, em que `resolve:ssm` é o prefixo padrão e *`parameter-ARN`* é o ARN do parâmetro do Systems Manager.

O exemplo a seguir cria um modelo de execução que obtém o ID de AMI de um parâmetro do Systems Manager existente com o ARN de *`arn:aws:ssm:us-east-2:123456789012:parameter/MyParameter`*.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling
 --version-description version1 \
  --launch-template-data file://config.json
```

Conteúdo de `config.json`:

```
{
    "ImageId":"resolve:ssm:arn:aws:ssm:us-east-2:123456789012:parameter/
MyParameter",
    "InstanceType":"t2.micro"
}
```

Quando nenhuma versão é especificada, a versão padrão do parâmetro é a versão mais recente.

O exemplo a seguir referencia uma versão específica do parâmetro *`MyParameter`*. O exemplo usa a versão *`3`* do parâmetro *`MyParameter`*, mas é possível usar qualquer número de versão válido.

```
{
    "ImageId":"resolve:ssm:arn:aws:ssm:us-east-2:123456789012:parameter/
MyParameter:3",
    "InstanceType":"t2.micro"
}
```

O exemplo semelhante a seguir referencia o rótulo de parâmetro *`prod`* que é mapeado em uma versão específica do parâmetro *`MyParameter`*.

```
{
    "ImageId":"resolve:ssm:arn:aws:ssm:us-east-2:123456789012:parameter/
MyParameter:prod",
    "InstanceType":"t2.micro"
}
```

O seguinte é um exemplo de resposta.

```
{
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-00f93d4588example",
        "LaunchTemplateName": "my-template-for-auto-scaling",
        "CreateTime": "2024-01-08T12:43:21.000Z",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
    }
}
```

Para especificar um parâmetro do Parameter Store em um modelo de execução, é necessário ter a permissão da `ssm:GetParameters` para o parâmetro especificado. Qualquer pessoa que use o modelo de execução também precisa da permissão da `ssm:GetParameters` para que o valor do parâmetro seja validado. Para obter exemplos de políticas do IAM, consulte Restringir o acesso a parâmetros do Systems Manager usando políticas do IAM no Guia do usuário do AWS Systems Manager .

Verifique se um modelo de execução obtém o ID da AMI correto

Use o describe-launch-template-versionscomando e inclua a `--resolve-alias` opção de resolver o parâmetro para a ID real da AMI.

```
aws ec2 describe-launch-template-versions --launch-template-name my-template-for-auto-
scaling \
   --versions 1 --resolve-alias
```

O exemplo retorna o ID de AMI para `ImageId`. Quando uma instância é iniciada usando esse modelo de execução, o ID de AMI é resolvido para `ami-0ac394d6a3example`.

```
{
```

```
    "LaunchTemplateVersions": [
        {
            "LaunchTemplateId": "lt-089c023a30example",
            "LaunchTemplateName": "my-template-for-auto-scaling",
            "VersionNumber": 1,
            "CreateTime": "2022-12-28T19:52:27.000Z",
            "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
            "DefaultVersion": true,
            "LaunchTemplateData": {
                "ImageId": "ami-0ac394d6a3example",
                "InstanceType": "t2.micro",
            }
        }
    ]
}
```

Recursos relacionados

Para mais detalhes sobre como especificar um parâmetro do Systems Manager no seu modelo de inicialização, consulte Usar um parâmetro do Systems Manager em vez de um ID de AMI no Guia do Usuário do Amazon EC2.

Para obter mais informações sobre como trabalhar com os parâmetros do Systems Manager, consulte os materiais de referência apresentados a seguir na documentação do Systems Manager.

- Para criar versões e rótulos de parâmetros, consulte Como trabalhar com versões de parâmetros e Trabalhar com rótulos de parâmetros
- Para obter informações sobre como pesquisar os parâmetros públicos da AMI com suporte para o Amazon EC2, consulte Calling AMI public parameters.
- Para obter informações sobre o compartilhamento de parâmetros com outras AWS contas ou por meio de AWS Organizations, consulte Como trabalhar com parâmetros compartilhados.
- Para obter informações sobre como monitorar se seus parâmetros foram criados com sucesso, consulte Suporte de parâmetros nativos para Amazon Machine Image IDs.

Limitações

Ao trabalhar com parâmetros do Systems Manager, observe as seguintes limitações:

- O Amazon EC2 Auto Scaling suporta somente a especificação IDs de AMI como parâmetros.

- Não há suporte para a criação ou atualização de grupos de instâncias mistas com seleção de tipo de instância baseada em atributos usando um modelo de execução que especifique um parâmetro do Systems Manager.
- Se o seu grupo do Auto Scaling usar um modelo de execução que especifique um parâmetro do Systems Manager, você não poderá iniciar uma atualização de instância com uma configuração desejada ou usando a correspondência de saltos.
- Se o seu grupo do Auto Scaling usar um modelo de execução que especifique um parâmetro do Systems Manager, não haverá suporte a grupos de alta atividade.
- Em cada chamada para criar ou atualizar seu grupo do Auto Scaling, o Amazon EC2 Auto Scaling resolverá o parâmetro do Systems Manager no modelo de execução. Se você usar parâmetros avançados ou limites de throughput mais altos, as chamadas frequentes ao Parameter Store (ou seja, a operação `GetParameters`) poderão aumentar os custos do Systems Manager, pois as cobranças são realizadas por interação com a API do Parameter Store. Para obter mais informações, consulte Preços do AWS Systems Manager.

Configurações de execução do Auto Scaling

Important

Limitações:

- Desde 1º de janeiro de 2023 não há mais suporte aos novos tipos de instância do Amazon EC2 nas configurações de execução. Isso inclui suporte para qualquer tipo de instância adicionado Região da AWS após o lançamento inicial da região.
- As contas criadas em ou após 1º de junho de 2023 não podem criar novas configurações de execução usando o console.
- As contas criadas em ou após 1º de outubro de 2024 não podem criar novas configurações de lançamento usando nenhum método (console AWS CLI, API ou CloudFormation).

Migre para modelos de execução para garantir que você não precise criar novas configurações de execução agora ou no futuro. Para obter informações sobre como migrar seu grupo do Auto Scaling, para lançar modelos, consulte Migre seus grupos do Auto Scaling para modelos de execução.

Note

Talvez você consiga criar uma configuração de execução com um tipo de instância que não tenha mais suporte em uma região. Recomendamos migrar para modelos de execução.

Fornecemos informações sobre configurações de execução para clientes que ainda não migraram das configurações de execução para os modelos de execução. Uma configuração de execução é um modelo de configuração de instância que um grupo do Auto Scaling usa para instâncias do EC2. Ao criar uma configuração de execução, você especifica informações para as instâncias. Inclua o ID da imagem de máquina da Amazon (AMI), o tipo de instância, um par de chaves, um ou mais grupos de segurança e um mapeamento de dispositivos de blocos. Se você tiver ativado uma instância do EC2 antes, você terá especificado as mesmas informações para ativar a instância.

Você pode especificar a configuração de execução com vários grupos do Auto Scaling. No entanto, você só pode especificar uma configuração de execução para um grupo do Auto Scaling de cada

vez, e você não pode modificar uma configuração de execução depois de criá-la. Para alterar a configuração de execução de um grupo do Auto Scaling, será necessário criar uma configuração de execução e, em seguida, atualizar seu grupo do Auto Scaling com ela.

Conteúdo

- Criar uma configuração de execução
- Alterar a configuração de execução de um grupo do Auto Scaling

Criar uma configuração de execução

Important

Limitações:

- Desde 1º de janeiro de 2023 não há mais suporte aos novos tipos de instância do Amazon EC2 nas configurações de execução. Isso inclui suporte para qualquer tipo de instância adicionado Região da AWS após o lançamento inicial da região.
- As contas criadas em ou após 1º de junho de 2023 não podem criar novas configurações de execução usando o console.
- As contas criadas em ou após 1º de outubro de 2024 não podem criar novas configurações de lançamento usando nenhum método (console AWS CLI, API ou CloudFormation).

Migre para modelos de execução para garantir que você não precise criar novas configurações de execução agora ou no futuro. Para obter informações sobre como migrar seu grupo do Auto Scaling, para lançar modelos, consulte Migre seus grupos do Auto Scaling para modelos de execução.

Note

Talvez você consiga criar uma configuração de execução com um tipo de instância que não tenha mais suporte em uma região. Recomendamos migrar para modelos de execução.

Este tópico descreve como criar uma configuração de inicialização. Fornecemos informações sobre configurações de execução para clientes que ainda não migraram das configurações de execução para os modelos de execução.

Não é possível modificar uma configuração de execução de uma instância após criá-la. Em vez disso, é necessário criar uma nova configuração de inicialização.

Para associar uma nova configuração de inicialização a um grupo do Auto Scaling existente, consulte Alterar a configuração de execução de um grupo do Auto Scaling. Para criar um grupo do Auto Scaling, consulte Criar um grupo do Auto Scaling usando uma configuração de execução

Conteúdo

- Criar uma configuração de execução
- Configurar as opções de metadados da instância
- Criar uma configuração de execução usando uma instância do EC2

Criar uma configuração de execução

Para criar uma configuração de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. Na barra de navegação superior, selecione sua AWS região.
3. No painel de navegação, em Ajuste de escala automático, escolha Grupos do Auto Scaling.
4. Escolha Configurações de execução na parte superior da página. Quando a confirmação for solicitada, escolha Exibir configurações de inicialização para confirmar que deseja visualizar a página Configurações de inicialização.
5. SelecioneCreate launch configuration (Criar uma configuração de execução), e insira um nome para sua configuração de execução.
6. Em Amazon machine image (AMI) (Imagem de máquina da Amazon (AMI)), escolha uma AMI. Para escolher uma AMI específica, você pode encontrar uma AMI adequada, anotar seu ID e inserir o ID como critério de pesquisa.

   Para obter a ID da AMI do Amazon Linux 2:

   a. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
   b. No painel de navegação esquerdo, em Instâncias, escolha Instâncias e, em seguida, escolha Iniciar instâncias.

c. Na guia Quick Start (Início rápido) da página Choose an Amazon Machine Image (Escolha uma Imagem de máquina da Amazon), observe o ID da AMI ao lado de Amazon Linux 2 AMI (HVM).

7. Na etapa Choose Instance Type (Escolher tipo de instância), selecione uma configuração de hardware para suas instâncias.
8. Em Additional configuration (Configuração adicional), preste atenção aos seguintes campos:

   a. (Opcional) Para Purchasing option (Opção de compra), você pode escolher Request Spot Instances (Solicitar instâncias spot) para solicitar instâncias spot ao preço spot, limitado ao preço sob demanda. Opcionalmente, você pode especificar um preço máximo por hora de instância para suas instâncias spot.

   > **Note**
   >
   > As instâncias spot são uma opção econômica em comparação com as instâncias sob demanda, se você puder ser flexível sobre quando suas aplicações são executadas e se for possível interromper suas aplicações. Para obter mais informações, consulte Solicitar instâncias spot para aplicações flexíveis e com tolerância a falhas.

   b. (Opcional) Em IAM instance profile (Perfil de instância do IAM) selecione um perfil a ser associado às instâncias. Para obter mais informações, consulte Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2.

   c. (Opcional) Para monitoramento, escolha se deseja permitir que as instâncias publiquem dados métricos em intervalos de 1 minuto na Amazon, CloudWatch ativando o monitoramento detalhado. Aplicam-se cobranças adicionais do . Para obter mais informações, consulte Configurar monitoramento para instâncias do Auto Scaling.

   d. (Opcional) Em Advanced details (Detalhes avançados), User data (Dados do usuário), você pode especificar dados do usuário para configurar uma instância durante a execução ou para executar um script de configuração após a instância ser iniciada.

   e. (Opcional) Em Advanced details (Detalhes avançados), IP address type (Tipo de endereço IP), escolha se deseja atribuir um public IP address (endereço IP público) às instâncias do grupo. Se você não definir um valor, o padrão é usar as configurações de IP público de atribuição automática das sub-redes nas quais suas instâncias são iniciadas.

9. (Opcional) Em Storage (volumes) (Armazenamento - volumes), se não precisar de armazenamento adicional, ignore esta seção. Caso contrário, para especificar os volumes a

serem anexados às instâncias, além dos volumes especificados pela AMI, escolha Add new volume (Adicionar novo volume). Em seguida, escolha as opções desejadas e os valores associados para Devices (Dispositivos), Snapshot, Size (Tamanho), Volume type (Tipo de volume), IOPS, Throughput (Taxa de transferência),Delete on termination (Excluir ao término), e Encrypted (Criptografado).

10. Em Security groups (Grupos de segurança), crie ou selecione o grupo de segurança para associar às instâncias do grupo. Se você mantiver a opção Create a new security group (Criar um novo grupo de segurança) selecionada, uma regra de SSH padrão será configurada para instâncias do Amazon EC2 que executem Linux. Uma função do RDP padrão é configurada para instâncias do Amazon EC2 que executem o Windows.
11. Em Key pair (login) (Par de chaves - login), escolha uma opção em Key pair options (Opções de par de chaves).

    Se já tiver configurado um par de chaves de instância do Amazon EC2, você pode escolhê-lo aqui.

    Caso você ainda não tenha um par de chaves da instância do Amazon EC2, escolha Create a new key pair (Criar um novo par de chaves) e atribua a ele um nome reconhecível. Escolha Download key pair (Fazer download do par de chaves) para fazer baixar o par de chaves para seu computador.

    > **Important**
    >
    > Não escolha Proceed without a key pair (Continuar sem um par de chaves) se você precisar se conectar à sua instância.

12. Selecione a caixa de confirmação e escolha Criar configuração de execução.

Para criar uma configuração de inicialização a partir de uma configuração de inicialização existente (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. Na barra de navegação superior, selecione sua AWS região.
3. No painel de navegação, em Ajuste de escala automático, escolha Grupos do Auto Scaling.
4. Escolha Configurações de execução na parte superior da página. Quando a confirmação for solicitada, escolha Exibir configurações de inicialização para confirmar que deseja visualizar a página Configurações de inicialização.

5. Selecione a configuração de execução e escolha Ações, Copiar configuração de execução. Isso configura uma nova configuração de execução com as mesmas opções da original, mas com "Copy" adicionado ao nome.
6. Na página Copiar configuração de execução, edite as opções de configuração conforme o necessário e escolha Criar configuração de execução.

Para criar uma configuração de execução usando a linha de comando

Você pode usar um dos comandos a seguir:

- create-launch-configuration (AWS CLI)
- Nova ASLaunch configuração (AWS Tools for Windows PowerShell)

Configurar as opções de metadados da instância

O Amazon EC2 Auto Scaling oferece suporte à configuração do Serviço de metadados da instância (IMDS) em configurações de execução. Isso lhe dá a opção de usar configurações de execução para configurar as instâncias do Amazon EC2 em seus grupos de Auto Scaling para exigir o Instance Metadata Service Version 2 IMDSv2 (), que é um método orientado por sessão para solicitar metadados de instâncias. Para obter detalhes sobre IMDSv2 as vantagens, consulte este artigo no AWS blog sobre aprimoramentos para adicionar defesa em profundidade ao serviço de metadados de instâncias do EC2.

Você pode configurar o IMDS para oferecer suporte a IMDSv2 e IMDSv1 (o padrão) ou para exigir o uso de IMDSv2. Se você estiver usando o AWS CLI ou um dos SDKs para configurar o IMDS, deverá usar a versão mais recente do AWS CLI ou do SDK para exigir o uso do. IMDSv2

Você pode configurar sua configuração de execução para:

- Exigir o uso de IMDSv2 ao solicitar metadados da instância
- Especificar o limite de salto de resposta PUT
- Desativar o acesso aos metadados da instância

Você pode encontrar mais detalhes sobre como configurar o Serviço de metadados da instância no tópico a seguir: Configuração do serviço de metadados da instância no Manual do usuário do Amazon EC2.

Use o seguinte procedimento para configurar as opções do IMDS em uma configuração de execução. Depois de criar sua configuração de execução, você pode associá-la ao seu grupo do Auto Scaling. Se você associar a configuração de execução a um grupo do Auto Scaling existente, a configuração de execução existente será desassociada do grupo do Auto Scaling e as instâncias existentes precisarão ser substituídas para usar as opções de IMDS especificadas na nova configuração de execução. Para obter mais informações, consulte Alterar a configuração de execução de um grupo do Auto Scaling.

Para configurar o IMDS em uma configuração de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. Na barra de navegação superior, selecione sua AWS região.
3. No painel de navegação, em Ajuste de escala automático, escolha Grupos do Auto Scaling.
4. Escolha Configurações de execução na parte superior da página. Quando a confirmação for solicitada, escolha Exibir configurações de inicialização para confirmar que deseja visualizar a página Configurações de inicialização.
5. Escolha Create launch configuration (Criar configuração de execução) e crie a configuração de execução da maneira usual. Inclua o ID da Imagem de máquina da Amazon (AMI), o tipo de instância e, opcionalmente, um par de chaves, um ou mais grupos de segurança e quaisquer volumes do EBS adicionais ou volumes de armazenamento de instâncias para suas instâncias.
6. Para configurar opções de metadados de instância para todas as instâncias associadas a esta configuração de execução, em Additional configuration (Configurações adicionais), em Advanced details (Detalhes avançados), faça o seguinte:

    a. Em Metadata accessible (Metadados acessíveis): escolha se deseja habilitar ou desabilitar o acesso ao endpoint do serviço de metadados da instância. Por padrão, o endpoint de HTTP está habilitado. Se você optar por desabilitar o endpoint, o acesso aos metadados da instância será desativado. Você pode especificar a condição a ser exigida IMDSv2 somente quando o endpoint HTTP estiver habilitado.

    b. Para a versão Metadata, você pode optar por exigir o uso do Instance Metadata Service Version 2 (IMDSv2) ao solicitar metadados da instância. Se você não especificar um valor, o padrão é oferecer suporte a ambos IMDSv1 IMDSv2 e.

    c. Em Metadata token response hop limit (Limite de salto de resposta do token de metadados), você pode definir o número permitido de saltos de rede para o token de metadados. Se você não especificar um valor, o padrão é 1.

7. Quando tiver concluído, escolha Create a launch configuration (Criar uma configuração de execução).

Para exigir o uso de IMDSv2 em uma configuração de inicialização usando o AWS CLI

Use o create-launch-configurationcomando a seguir com `--metadata-options` definido como`HttpTokens=required`. Quando você especifica um valor para `HttpTokens`, você também deve definir `HttpEndpoint` como ativado. Como o cabeçalho do token seguro está definido como obrigatório para solicitações de recuperação de metadados, isso faz com que a instância exija o uso IMDSv2 ao solicitar metadados da instância.

```
aws autoscaling create-launch-configuration \
  --launch-configuration-name my-lc-with-imdsv2 \
  --image-id ami-01e24be29428c15b2 \
  --instance-type t2.micro \
      ...
  --metadata-options "HttpEndpoint=enabled,HttpTokens=required"
```

Como desabilitar o acesso aos metadados da instância

Use o create-launch-configurationcomando a seguir para desativar o acesso aos metadados da instância. Você pode reativar o acesso posteriormente usando o modify-instance-metadata-optionscomando.

```
aws autoscaling create-launch-configuration \
  --launch-configuration-name my-lc-with-imds-disabled \
  --image-id ami-01e24be29428c15b2 \
  --instance-type t2.micro \
      ...
  --metadata-options "HttpEndpoint=disabled"
```

Criar uma configuração de execução usando uma instância do EC2

A opção para criar uma configuração de execução usando os atributos de uma instância do EC2 em execução está disponível.

Há diferenças entre a criação de uma configuração de execução do zero e a criação de uma configuração de execução a partir de uma instância do EC2. Quando você cria uma configuração de execução do zero, você especifica o ID de imagem, o tipo de instância, os recursos opcionais

(como dispositivos de armazenamento) e configurações opcionais (como monitoramento). Quando você cria uma configuração de execução a partir de uma instância em execução, o Amazon EC2 Auto Scaling gera atributos para a configuração de execução a partir da instância especificada. Os atributos são também derivados do mapeamento de dispositivos de blocos para a AMI da qual a instância foi executada, ignorando todos os outros dispositivos de blocos que foram adicionados após a execução.

Ao criar uma configuração de execução usando uma instância em execução, você pode substituir os atributos a seguir especificando-os como parte da mesma solicitação: AMI, dispositivos de blocos, par de chaves, perfil de instância, tipo de instância, kernel, monitoramento de instância, locação de localização, ramdisk, grupos de segurança, preço spot (máximo), dados do usuário, se a instância tem um endereço IP público e se a instância é otimizada para EBS.

> **Note**
>
> Se a instância especificada tiver propriedades que atualmente não são suportadas pelas configurações de execução, as instâncias executadas pelo grupo do Auto Scaling podem não ser idênticas à instância original do EC2.

> **Important**
>
> A AMI usada para ativar a instância especificada ainda deve existir.

Tópicos

- Criar uma configuração de execução a partir de uma instância do EC2 (AWS CLI)
- Criar uma configuração de execução a partir de uma instância e substituir os dispositivos de blocos (AWS CLI)
- Criar uma configuração de execução e substituir o tipo de instância (AWS CLI)

Criar uma configuração de execução a partir de uma instância do EC2 (AWS CLI)

Use o comando create-launch-configuration a seguir para criar uma configuração de execução a partir de uma instância usando os mesmos atributos que a instância. Todos os dispositivos de blocos adicionados após a execução são ignorados.

```
aws autoscaling create-launch-configuration --launch-configuration-name my-lc-from-instance --instance-id i-a8e09d9c
```

Você pode usar o describe-launch-configurationscomando a seguir para descrever a configuração de execução e verificar se seus atributos correspondem aos da instância.

```
aws autoscaling describe-launch-configurations --launch-configuration-names my-lc-from-instance
```

O seguinte é um exemplo de resposta.

```
{
    "LaunchConfigurations": [
        {
            "UserData": null,
            "EbsOptimized": false,
            "LaunchConfigurationARN": "arn",
            "InstanceMonitoring": {
                "Enabled": false
            },
            "ImageId": "ami-05355a6c",
            "CreatedTime": "2014-12-29T16:14:50.382Z",
            "BlockDeviceMappings": [],
            "KeyName": "my-key-pair",
            "SecurityGroups": [
                "sg-8422d1eb"
            ],
            "LaunchConfigurationName": "my-lc-from-instance",
            "KernelId": "null",
            "RamdiskId": null,
            "InstanceType": "t1.micro",
            "AssociatePublicIpAddress": true
        }
    ]
}
```

Criar uma configuração de execução a partir de uma instância e substituir os dispositivos de blocos (AWS CLI)

Por padrão, o Amazon EC2 Auto Scaling usa os atributos da instância do EC2 que você especifica para criar a configuração de execução. No entanto, os dispositivos de blocos são provenientes

da AMI usada para iniciar a instância, não a instância. Para adicionar dispositivos de blocos à configuração de execução, substitua o mapeamento de dispositivos de blocos para a configuração de execução.

Use o create-launch-configurationcomando a seguir para criar uma configuração de execução usando uma instância do EC2, mas com um mapeamento personalizado de dispositivos de blocos.

```
aws autoscaling create-launch-configuration --launch-configuration-name my-lc-from-instance-bdm --instance-id i-a8e09d9c \
  --block-device-mappings "[{\"DeviceName\":\"/dev/sda1\",\"Ebs\":{\"SnapshotId\":\"snap-3decf207\"}},{\"DeviceName\":\"/dev/sdf\",\"Ebs\":{\"SnapshotId\":\"snap-eed6ac86\"}}]"
```

Use o describe-launch-configurationscomando a seguir para descrever a configuração de inicialização e verificar se ela usa seu mapeamento personalizado de dispositivos de blocos.

```
aws autoscaling describe-launch-configurations --launch-configuration-names my-lc-from-instance-bdm
```

A resposta do exemplo a seguir descreve a configuração de execução.

```
{
    "LaunchConfigurations": [
        {
            "UserData": null,
            "EbsOptimized": false,
            "LaunchConfigurationARN": "arn",
            "InstanceMonitoring": {
                "Enabled": false
            },
            "ImageId": "ami-c49c0dac",
            "CreatedTime": "2015-01-07T14:51:26.065Z",
            "BlockDeviceMappings": [
                {
                    "DeviceName": "/dev/sda1",
                    "Ebs": {
                        "SnapshotId": "snap-3decf207"
                    }
                },
                {
                    "DeviceName": "/dev/sdf",
                    "Ebs": {
```

```
                    "SnapshotId": "snap-eed6ac86"
                }
            }
        ],
        "KeyName": "my-key-pair",
        "SecurityGroups": [
            "sg-8637d3e3"
        ],
        "LaunchConfigurationName": "my-lc-from-instance-bdm",
        "KernelId": null,
        "RamdiskId": null,
        "InstanceType": "t1.micro",
        "AssociatePublicIpAddress": true
    }
  ]
}
```

Criar uma configuração de execução e substituir o tipo de instância (AWS CLI)

Por padrão, o Amazon EC2 Auto Scaling usa os atributos da instância do EC2 que você especifica para criar a configuração de execução. Dependendo dos seus requisitos, convém substituir atributos da instância e usar os valores que você precisa. Por exemplo, você pode substituir o tipo de instância.

Use o create-launch-configurationcomando a seguir para criar uma configuração de execução usando uma instância do EC2, mas com um tipo de instância diferente (por exemplo`t2.medium`) do que a instância (por exemplo`t2.micro`).

```
aws autoscaling create-launch-configuration --launch-configuration-name my-lc-from-instance-changetype \
  --instance-id i-a8e09d9c --instance-type t2.medium
```

Use o describe-launch-configurationscomando a seguir para descrever a configuração de execução e verificar se o tipo de instância foi substituído.

```
aws autoscaling describe-launch-configurations --launch-configuration-names my-lc-from-instance-changetype
```

A resposta do exemplo a seguir descreve a configuração de execução.

```
{
```

```
    "LaunchConfigurations": [
        {
            "UserData": null,
            "EbsOptimized": false,
            "LaunchConfigurationARN": "arn",
            "InstanceMonitoring": {
                "Enabled": false
            },
            "ImageId": "ami-05355a6c",
            "CreatedTime": "2014-12-29T16:14:50.382Z",
            "BlockDeviceMappings": [],
            "KeyName": "my-key-pair",
            "SecurityGroups": [
                "sg-8422d1eb"
            ],
            "LaunchConfigurationName": "my-lc-from-instance-changetype",
            "KernelId": "null",
            "RamdiskId": null,
            "InstanceType": "t2.medium",
            "AssociatePublicIpAddress": true
        }
    ]
}
```

Alterar a configuração de execução de um grupo do Auto Scaling

> **Important**
>
> Fornecemos informações sobre configurações de execução para clientes que ainda não migraram das configurações de execução para os modelos de execução. Para obter informações sobre como migrar seu grupo do Auto Scaling, para lançar modelos, consulte Migre seus grupos do Auto Scaling para modelos de execução.

Este tópico descreve como associar uma configuração de inicialização diferente ao seu grupo do Auto Scaling.

Depois que você alterar a configuração de execução todas as novas instâncias serão ativadas usando as novas opções de configuração, mas as instâncias existentes não serão afetadas. Para obter mais informações, consulte Atualizar instâncias do Auto Scaling.

Para alterar a configuração de execução para um grupo do Auto Scaling (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação esquerdo, em Auto Scaling, escolha Grupos do Auto Scaling.
3. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
4. Na guia Details (Detalhes), escolha Launch configuration (Configuração de execução), Edit (Editar).
5. Em Launch Configuration (Configuração de execução), selecione a nova configuração de execução.
6. Quando terminar, escolha Update (Atualizar).

Para alterar a configuração de execução para um grupo do Auto Scaling (console) usando a linha de comando.

Você pode usar um dos comandos a seguir:

- update-auto-scaling-group (AWS CLI)
- Atualizar- ASAuto ScalingGroup (AWS Tools for Windows PowerShell)

Grupos do Auto Scaling

> **Note**
>
> Se você não tem experiência com grupos do Auto Scaling, siga as etapas do tutorial Criar seu primeiro grupo do Auto Scaling para começar e ver como um grupo do Auto Scaling responde quando uma instância do grupo é encerrada.

Um grupo do Auto Scaling contém um conjunto de instâncias do EC2 que são tratadas como um agrupamento lógico para fins de gerenciamento e escalabilidade automática. Um grupo do Auto Scaling também permite que você use recursos do Amazon EC2 Auto Scaling como substituições de verificação de integridade e políticas de escalabilidade. A manutenção do número de instâncias em um grupo do Auto Scaling e a escalabilidade automática são os principais recursos do serviço Amazon EC2 Auto Scaling.

O tamanho de um grupo do Auto Scaling depende do número de instâncias definidas como a capacidade desejada. Você pode ajustar seu tamanho para atender à demanda, manualmente ou usando a escalabilidade automática.

Um grupo do Auto Scaling começa iniciando instâncias suficientes para atender à sua capacidade desejada. Ele mantém esse número de instâncias executando verificações de integridade periódicas nas instâncias do grupo. O grupo do Auto Scaling continua a manter um número fixo de instâncias, mesmo que uma instância se torne não íntegra. Se uma instância se tornar não íntegra, o grupo a encerrará e iniciará outra instância para substituí-la. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

É possível usar políticas de escalabilidade para aumentar ou diminuir o número de instâncias em seu grupo dinamicamente para atender a condições em alteração. Quando a política de escalabilidade está habilitada, o grupo do Auto Scaling ajusta a capacidade desejada do grupo, entre os valores mínimo e máximo de capacidade especificados, e inicia ou termina as instâncias, conforme necessário. Você também pode dimensionar com base em uma programação. Para obter mais informações, consulte Escolha seu método de escalabilidade.

Ao criar um grupo do Auto Scaling você pode optar por iniciar instâncias sob demanda, instâncias spot ou ambas. Você pode especificar várias opções de compra para seu grupo do Auto Scaling somente quando você usa um launch template Para obter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

As instâncias spot permitem que você acesse a capacidade não utilizada do EC2 com grandes descontos em relação aos preços sob demanda. Para obter mais informações, consulte Instâncias spot do Amazon EC2. Existem diferenças importantes entre instâncias spot e instâncias sob demanda:

- O preço das instâncias spot varia de acordo com a demanda
- O Amazon EC2 pode terminar uma Instância spot individual conforme a disponibilidade ou o preço das instâncias spot for alterado

Quando uma instância spot é terminada, o grupo do Auto Scaling tenta iniciar uma instância de substituição para manter a capacidade desejada para o grupo.

Quando as instâncias são executadas, se você especificou várias zonas de disponibilidade, a capacidade desejada é distribuída entre essas zonas de disponibilidade. Se ocorrer uma ação de escalabilidade, o Amazon EC2 Auto Scaling manterá automaticamente o equilíbrio entre todas as zonas de disponibilidade especificadas.

Conteúdo

- Criar grupos do Auto Scaling usando modelos de execução
- Criar grupos do Auto Scaling usando configurações de execução
- Execute instâncias de forma síncrona
- Update an Auto Scaling group
- Etiquetar grupos e instâncias do Auto Scaling
- Políticas de manutenção de instância
- Ganchos do ciclo de vida do Amazon EC2 Auto Scaling
- Diminua a latência para aplicações com tempos de inicialização longos usando pools ativos
- Deslocamento zonal do grupo Auto Scaling
- Distribuição de zonas de disponibilidade de grupo do Auto Scaling
- Desvincular instâncias do seu grupo do Auto Scaling
- Remover temporiamente instâncias do grupo do Auto Scaling
- Excluir infraestrutura do Auto Scaling

Criar grupos do Auto Scaling usando modelos de execução

Antes de criar um modelo de execução, você pode criar um grupo do Auto Scaling que use um modelo de execução como modelo de configuração para as instâncias do EC2. O modelo de execução especifica informações, como o ID da AMI, o tipo de instância, o par de chaves, os grupos de segurança e o mapeamento de dispositivos de blocos para suas instâncias. Para obter mais informações sobre como criar modelos de execução, consulte Criar um modelo de execução para um grupo do Auto Scaling.

É necessário ter permissões suficientes para criar um grupo do Auto Scaling. Você também deve ter permissões suficientes para criar o perfil vinculado ao serviço que o Amazon EC2 Auto Scaling usa para realizar ações por sua própria conta se ela ainda não existir. Para exemplos de políticas do IAM que um administrador pode usar como referência para conceder permissões a você, consulte Exemplos de políticas baseadas em identidade e Controle o uso do modelo de execução do Amazon EC2 em grupos do Auto Scaling.

Conteúdo

- Criar um grupo do Auto Scaling usando um modelo de execução
- Criar um grupo do Auto Scaling usando o assistente de execução do Amazon EC2
- Grupos do Auto Scaling com vários tipos de instâncias e opções de compra

Criar um grupo do Auto Scaling usando um modelo de execução

Ao criar um grupo do Auto Scaling, será necessário especificar as informações necessárias para configurar as instâncias do Amazon EC2, as zonas de disponibilidade e sub-redes VPC para as instâncias, a capacidade desejada e os limites de capacidade mínimo e máximo.

Para configurar instâncias do Amazon EC2 que são executadas pelo seu grupo do Auto Scaling, é possível especificar um modelo de execução ou uma configuração de execução. O procedimento a seguir demonstra como criar um grupo do Auto Scaling usando um modelo de execução.

Pré-requisitos

- É necessário ter criado um modelo de execução. Para obter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.

Para criar um grupo do Auto Scaling usando um modelo de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a mesma Região da AWS que você usou ao criar o modelo de lançamento.
3. Selecione Criar um grupo do Auto Scaling.
4. Na página Choose launch template or configuration (Escolher modelo de execução ou configuração) faça o seguinte:

   a. Em Auto Scaling group name (Nome do grupo do Auto Scaling), insira um nome para o seu grupo do Auto Scaling.

   b. Em Launch template (Modelo de execução), escolha um modelo de execução existente.

   c. Em Launch template version (Versão do modelo de execução), indique se o grupo do Auto Scaling usará a versão padrão, a mais recente ou uma versão específica do modelo de execução no aumento da escala na horizontal.

   d. Verifique se o modelo de execução oferece suporte a todas as opções que você está planejando usar e escolha Next (Próximo).
5. Na página Opções de iniciação de escolha, se você não estiver usando vários tipos de instância, pode pular a seção Requisitos de tipo de instância para usar o tipo de instância do EC2 especificado no modelo de lançamento.

   Para usar vários tipos de instância, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.
6. Em Network (Rede), para VPC, escolha uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução.
7. Em Zonas de disponibilidade e sub-redes, escolha uma ou mais sub-redes na VPC especificada. Use sub-redes em várias zonas de disponibilidade para alta disponibilidade. Para obter mais informações, consulte Considerações sobre a escolha de sub-redes da VPC.
8. Em Distribuição na zona de disponibilidade, selecione uma estratégia de distribuição. Para obter mais informações, consulte Distribuição de zonas de disponibilidade de grupo do Auto Scaling.
9. Se você criou um modelo de execução com um tipo de instância especificado, poderá continuar para a próxima etapa para criar um grupo do Auto Scaling que use o tipo de instância no modelo de execução.

Como alternativa, você pode escolher Override launch template (Substituir modelo de execução) se nenhum tipo de instância for especificado no modelo de execução ou se você quiser usar vários tipos de instância para autoescalabilidade. Para obter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

10. Selecione Next (Próximo) para continuar para a próxima etapa.

    Ou é possível aceitar o restante dos padrões e escolher Skip to review (Avançar para análise).

11. (Opcional) Na página Integrar com outros serviços, configure as opções a seguir e escolha Próximo:

    a. Em Balanceamento de carga, escolha se deseja vincular seu grupo do Auto Scaling a um balanceador de carga. Para obter mais informações, consulte Elastic Load Balancing.

    b. Em Opções de integração com o VPC Lattice, escolha se deseja usar o VPC Lattice. Para obter mais informações, consulte Gerenciar o fluxo de tráfego com um grupo de destino do VPC Lattice.

    c. Em Mudança de zona do Controlador de Recuperação de Aplicações (ARC) da Amazon, selecione a caixa de seleção para habilitar mudança de zona. Para obter mais informações, consulte Deslocamento zonal do grupo Auto Scaling.

       - Se você habilitar mudança de zona, em Comportamento da verificação de integridade, selecione Ignorar não íntegro ou Substituir não íntegro. Para obter mais informações, consulte Como a mudança zonal funciona para grupos de Auto Scaling.

    d. Em Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Amazon EBS. Para obter mais informações, consulte Monitore as instâncias do Auto Scaling com volumes do Amazon EBS com problemas usando verificações de integridade.

    e. Em Período de carência da verificação de integridade, insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

12. (Opcional) Na página Configurar o tamanho do grupo e escalabilidade, configure as opções a seguir e escolha Próximo:

    a. Sob o tamanho do grupo, para a capacidade desejada, insira o número inicial de instâncias para ser lançado.

b. Em Escalabilidade, Limites de escalabilidade, se o novo valor para Capacidade desejada for maior que a Capacidade mínima desejada e Capacidade máxima desejada, a Capacidade máxima desejada será automaticamente aumentada para o novo valor da capacidade desejada. É possível alterar esses limites conforme necessário. Para obter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.

c. Em Escalabilidade automática, escolha se você deseja criar uma política de escalabilidade de rastreamento de destino. Você também pode criar essa política depois de criar seu grupo do Auto Scaling.

   Se você escolher a política de escalabilidade de rastreamento de destino, siga as instruções em Criar uma política de dimensionamento com monitoramento do objetivo para criar a política.

d. Em Política de manutenção de instâncias, escolha se você deseja criar uma política de manutenção de instâncias. Você também pode criar essa política depois de criar seu grupo do Auto Scaling. Siga as instruções Definir uma política de manutenção de instâncias para criar a política.

e. Em Configurações adicionais de capacidade, Preferência de reserva de capacidade, escolha se deseja usar uma preferência de reserva de capacidade. Para obter mais informações, consulte Reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas.

f. Em Configurações adicionais, Proteção contra redução de escala na horizontal de instâncias, escolha se deseja habilitar a proteção contra redução de escala na horizontal de instâncias. Para obter mais informações, consulte Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância.

g. Em Monitoramento, escolha se deseja ativar a coleta de métricas de CloudWatch grupo. Essas métricas fornecem medições que podem ser indicadores de um problema potencial, como número de instâncias de terminação ou número de instâncias pendentes. Para obter mais informações, consulte Monitore CloudWatch métricas para seus grupos e instâncias do Auto Scaling.

h. Em Aquecimento de instância padrão, selecione essa opção e escolha o tempo de aquecimento para sua aplicação. Se você estiver criando um grupo de Auto Scaling que tenha uma política de escalabilidade, o recurso padrão de aquecimento de instâncias melhora as CloudWatch métricas da Amazon usadas para escalabilidade dinâmica. Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

13. (Opcional) Na página Adicionar notificações, configure a notificação e, em seguida, escolha Próximo. Para obter mais informações, consulte Configurar notificações do Amazon SNS para o Amazon EC2 Auto Scaling.
14. (Opcional) Na página Adicionar tags, escolha Adicionar tag, forneça uma chave e um valor para cada tag e, em seguida, escolha Próximo. Para obter mais informações, consulte Etiquetar grupos e instâncias do Auto Scaling.
15. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Para criar um grupo do Auto Scaling usando a linha de comando

Você pode usar um dos comandos a seguir:

- create-auto-scaling-group (AWS CLI)
- Novo- ASAuto ScalingGroup (AWS Tools for Windows PowerShell)

Criar um grupo do Auto Scaling usando o assistente de execução do Amazon EC2

O procedimento a seguir mostra como criar um grupo do Auto Scaling usando o assistente Launch instance (Iniciar instância) no console do Amazon EC2. Essa opção preenche automaticamente o modelo de execução com determinados detalhes de configuração do assistente Launch instance (Iniciar instância).

> **Note**
>
> O assistente não preenche o grupo do Auto Scaling com o número de instâncias especificadas; ele só preenche o modelo de execução com o ID e o tipo de instância da imagem de máquina da Amazon (AMI). Usar o assistente Create Auto Scaling group (Criar grupo do Auto Scaling) para especificar o número de instâncias a serem iniciadas.
> Uma AMI fornece as informações necessárias para configurar uma instância. É possível iniciar várias instâncias em uma única AMI quando precisa de várias instâncias com a mesma configuração. Recomendamos usar uma AMI personalizada que já tenha sua aplicação instalada nela para evitar que suas instâncias sejam terminadas se você reiniciar uma instância pertencente a um grupo do Auto Scaling. Para usar uma AMI personalizada com o Amazon EC2 Auto Scaling, é necessário primeiro criar sua AMI a partir de uma

instância personalizada e, em seguida, usar a AMI para criar um modelo de execução para o grupo do Auto Scaling.

Pré-requisitos

- Você deve ter criado uma AMI personalizada na mesma área Região da AWS em que planeja criar o grupo Auto Scaling. Para obter mais informações, consulte Criar uma AMI no Guia do usuário do Amazon EC2.

Use uma AMI personalizada como modelo

Nesta seção, você usa o assistente de execução do Amazon EC2 para preencher automaticamente um modelo de execução com a AMI personalizada. Como alternativa, para configurar o modelo de execução do zero ou para obter mais descrição dos parâmetros que você pode configurar para o seu modelo de execução, consulte Criar seu modelo de execução (console).

Para usar uma AMI personalizada como um modelo

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. Na barra de navegação na parte superior da tela, a corrente Região da AWS é exibida. Selecione uma região na qual iniciará o grupo do Auto Scaling.
3. No painel de navegação, escolha Instâncias.
4. Escolha Launch instance (Iniciar instância) e faça o seguinte:
    a. Em Name and tags (Nome e etiquetas), deixe Name (Nome) em branco. O nome não faz parte dos dados usados para criar um modelo de execução.
    b. Em Imagens do aplicativo e do sistema operacional (Amazon Machine Image), escolha Procurar mais AMIs para ver o catálogo completo da AMI.
    c. Escolha Meu AMIs, encontre a AMI que você criou e escolha Selecionar.
    d. Em Instance type (Tipo de instância), escolha um tipo de instância.

> **Note**
>
> Escolha o mesmo tipo de instância que você usou quando criou a AMI ou uma mais potente.

e. No lado direito da tela, em Summary (Resumo), para Number of instances (Número de instâncias), insira qualquer número. O número que você insere aqui não é importante. Você especificará o número de instâncias que quer iniciar ao criar o grupo do Auto Scaling.

   No campo Number of instances (Número de instâncias), é exibida a mensagem When launching more than 1 instance, consider EC2 Auto Scaling (Ao iniciar mais de uma instância, considere o EC2 Auto Scaling).

f. Escolha o texto de hiperlink consider EC2 Auto Scaling (considerar o EC2 Auto Scaling).

g. No diálogo de confirmação Launch into Auto Scaling Group (Iniciar no grupo do Auto Scaling), escolha Continue (Continuar) para ir até a página Create launch template (Criar modelo de execução) com a AMI e o tipo de instância que você selecionou no assistente de instância de execução já preenchido.

Depois de escolher Continuar, a página Create launch template (Criar modelo de execução) é aberta. Siga este procedimento para concluir a criação de um modelo de execução.

Para criar um modelo de execução

1. Em Launch template name and description (Nome e descrição do modelo de execução), insira um nome e uma descrição para o modelo de execução.
2. (Opcional) Em Key pair (login) (Par de chaves [login]), Key pair name (Nome do par de chaves), escolha o nome do par de chaves criado anteriormente a ser usado quando você se conectar às instâncias, por exemplo, usando SSH.
3. (Opcional) Em Network settings (Configurações de rede), em Security groups (Grupos de segurança), escolha um ou mais grupos de segurança criados previamente.
4. (Opcional) Em Configure storage (Configurar armazenamento), atualize a configuração de armazenamento. A configuração de armazenamento padrão é determinada pela AMI e pelo tipo de instância.
5. Quando terminar de configurar o modelo de execução, selecione Create launch template (Criar modelo de execução).
6. Na página de confirmação, escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Criar um grupo do Auto Scaling

> **Note**
>
> O restante deste tópico descreve o procedimento básico para a criação de um grupo do Auto Scaling. Para obter mais descrição dos parâmetros que você pode configurar para o seu grupo do Auto Scaling, consulte Criar um grupo do Auto Scaling usando um modelo de execução.

Depois de escolher Create Auto Scaling group (Criar grupo do Auto Scaling), o assistente Create Auto Scaling group (Criar grupo do Auto Scaling) é aberto. Siga este procedimento para criar um grupo do Auto Scaling.

Para criar um grupo do Auto Scaling

1. Na página Choose launch template or configuration (Escolher o modelo ou a configuração de execução), insira um nome para o grupo de Auto Scaling.
2. O modelo de execução que você criou já está selecionado para você.

   Em Launch template version (Versão do modelo de execução), indique se o grupo do Auto Scaling usará a versão padrão, a mais recente ou uma versão específica do modelo de execução no aumento da escala na horizontal.
3. Selecione Next (Próximo) para continuar para a próxima etapa.
4. Na página Opções de iniciação de escolha, se você não estiver usando vários tipos de instância, pode pular a seção Requisitos de tipo de instância para usar o tipo de instância do EC2 especificado no modelo de lançamento.

   Para usar vários tipos de instância, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.
5. Em Network (Rede), para VPC, escolha uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução.

   > **Tip**
   >
   > Se você não especificou um grupo de segurança no modelo de execução, suas instâncias serão executadas com um grupo de segurança padrão da VPC que você

especificar. Por padrão, esse grupo de segurança não permite tráfego de entrada de redes externas.

6. Em Zonas de disponibilidade e sub-redes, escolha uma ou mais sub-redes na VPC especificada.
7. Em Distribuição na zona de disponibilidade, selecione uma estratégia de distribuição. Para obter mais informações, consulte Distribuição de zonas de disponibilidade de grupo do Auto Scaling.
8. Selecione Próximo duas vezes para ir até a página Configure group size and scaling policies (Definir tamanho do grupo e políticas de escalabilidade).
9. Em Tamanho do grupo, defina a Capacidade desejada (número inicial de instâncias para iniciar imediatamente após a criação do grupo do Auto Scaling).
10. Na seção Escalabilidade, em Limites de escalabilidade, se o novo valor para Capacidade desejada for maior que Capacidade mínima desejada e Capacidade máxima desejada, a Capacidade máxima desejada será automaticamente aumentada para o novo valor da capacidade desejada. É possível alterar esses limites conforme necessário. Para obter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.
11. Escolha Skip to review (Ir para revisão).
12. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Próximas etapas

Você pode conferir se o grupo do Auto Scaling foi criado corretamente visualizando o histórico de atividades. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status exibe se o seu grupo do Auto Scaling lançou instâncias com êxito. Se as instâncias não forem executadas ou forem executadas, mas terminadas imediatamente, consulte os tópicos a seguir para possíveis causas e resoluções:

- Solucionar problemas do Amazon EC2 Auto Scaling: falhas ao iniciar instâncias do EC2
- Solucionar problemas do Amazon EC2 Auto Scaling: problemas de AMI
- Solução de problemas de instâncias não íntegras no Amazon EC2 Auto Scaling

Agora você pode anexar um balanceador de carga na mesma região do grupo do Auto Scaling, se desejar. Para obter mais informações, consulte Use o Elastic Load Balancing para distribuir automaticamente o tráfego de entrada de aplicações no seu grupo do Auto Scaling..

Grupos do Auto Scaling com vários tipos de instâncias e opções de compra

Você pode iniciar e escalar automaticamente uma frota de instâncias sob demanda e instâncias spot em um único grupo do Auto Scaling. Além de receber descontos pelo uso de instâncias spot, é possível usar instâncias reservadas ou um Savings Plans para receber descontos no preço normal de instância sob demanda. Esses fatores ajudam você a otimizar sua economia de custos para instâncias do EC2 e a obter a escalabilidade e o desempenho desejados para sua aplicação.

As instâncias spot são capacidade de reserva disponível com grandes descontos em comparação com o preço do EC2 On-Demand. As Instâncias spot são uma opção econômica se houver flexibilidade quanto ao momento em que as aplicações serão executadas e se as aplicações poderão ser interrompidas. Você pode usar Instâncias spot para várias aplicações flexíveis e tolerantes a falhas. Os exemplos incluem servidores web sem estado, endpoints de API, aplicativos de big data e análise, cargas de trabalho em contêineres, CI/CD pipelines, computação de alto desempenho e alto rendimento (HPC/HTC), cargas de trabalho de renderização e outras cargas de trabalho flexíveis.

Para obter mais informações sobre as opções de compra, consulte Opções de compra de instância no Guia do usuário do Amazon EC2.

Tópicos

- Visão geral da configuração para criar um grupo de instâncias mistas
- Estratégias de alocação para vários tipos de instância
- Crie um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos
- Criar um grupo misto de instâncias escolhendo manualmente os tipos de instância
- Configurar um grupo do Auto Scaling para usar ponderação da instância
- Usar vários modelos de execução

Visão geral da configuração para criar um grupo de instâncias mistas

Este tópico fornece uma visão geral e as melhores práticas para criar um grupo de instâncias mistas.

Conteúdo

- Visão geral do
- Flexibilidade de tipo da instância
- Flexibilidade da zona de disponibilidade

- Preço máximo do spot
- Rebalanceamento proativo de capacidade
- Comportamento do ajuste de escala
- Disponibilidade regional dos tipos de instância
- Recursos relacionados
- Limitações

Visão geral do

Para criar um grupo de instâncias mistas, você tem duas opções:

- Seleção de tipo de instância com base em atributos — Defina seus requisitos de computação para escolher seus tipos de instância automaticamente com base em seus atributos de instância específicos.
- Seleção manual do tipo de instância — Escolha manualmente os tipos de instância adequados ao seu workload.

Manual selection

As etapas a seguir descrevem como criar um grupo de instâncias mistas escolhendo manualmente os tipos de instância:

1. Escolha um modelo de execução que tenha os parâmetros para iniciar uma instância do EC2. Os parâmetros nos modelos de execução são opcionais, mas o Amazon EC2 Auto Scaling não pode iniciar uma instância se o ID da amilong; (AMI) estiver ausente do modelo de execução.
2. Escolha a opção de substituir o modelo de execução.
3. Escolha manualmente os tipos de instância adequados ao seu workload.
4. Especifique as porcentagens de instâncias sob demanda e de instâncias spot a serem iniciadas.
5. Escolha estratégias de alocação que determinem como o Amazon EC2 Auto Scaling atenderá à capacidade sob demanda e spot com os tipos de instância possíveis.
6. Escolha as zonas de disponibilidade e sub-redes VPC nas quais iniciar suas instâncias.
7. Especifique o tamanho inicial do grupo (a capacidade desejada) e o tamanho mínimo e máximo do grupo.

As substituições são necessárias para substituir o tipo de instância declarado no modelo de execução e usar vários tipos de instâncias incorporados na própria definição de recursos do grupo do Auto Scaling. Para obter mais informações sobre os tipos de instâncias disponíveis, consulte Tipos de instância no Manual do usuário do Amazon EC2 para instâncias do Linux.

Você também pode configurar os seguintes parâmetros opcionais para cada tipo de instância:

- `LaunchTemplateSpecification` — Você pode atribuir um modelo de execução diferente a um tipo de instância, conforme necessário. Essa opção não está disponível atualmente no console. Para obter mais informações, consulte Usar vários modelos de execução.
- `WeightedCapacity` — Você decide o quanto a instância conta para a capacidade desejada em relação ao resto das instâncias do seu grupo. Se você especificar um valor `WeightedCapacity` para um tipo de instância, deverá especificar um valor `WeightedCapacity` para todos os tipos. Por padrão, cada instância conta como uma para a capacidade desejada. Para obter mais informações, consulte Configurar um grupo do Auto Scaling para usar ponderação da instância.

Attribute-based selection

Para permitir que o Amazon EC2 Auto Scaling escolha seus tipos de instância automaticamente com base em seus atributos de instância específicos, use as seguintes etapas para criar um grupo misto de instâncias especificando seus requisitos computacionais:

1. Escolha um modelo de execução que tenha os parâmetros para iniciar uma instância do EC2. Os parâmetros nos modelos de execução são opcionais, mas o Amazon EC2 Auto Scaling não pode iniciar uma instância se o ID da amilong; (AMI) estiver ausente do modelo de execução.
2. Escolha a opção de substituir o modelo de execução.
3. Especifique os atributos da instância que correspondam aos seus requisitos de computação, como v CPUs e requisitos de memória.
4. Especifique as porcentagens de instâncias sob demanda e de instâncias spot a serem iniciadas.
5. Escolha estratégias de alocação que determinem como o Amazon EC2 Auto Scaling atenderá à capacidade sob demanda e spot com os tipos de instância possíveis.
6. Escolha as zonas de disponibilidade e sub-redes VPC nas quais iniciar suas instâncias.
7. Especifique o tamanho inicial do grupo (a capacidade desejada) e o tamanho mínimo e máximo do grupo.

As substituições são necessárias para substituir o tipo de instância declarado no modelo de execução e usar um conjunto de atributos de instância que descrevam seus requisitos de computação. Para ver os atributos compatíveis, consulte a InstanceRequirementsReferência da API Amazon EC2 Auto Scaling. Como alternativa, é possível usar um modelo de execução que já tenha sua definição de atributos de instância.

Você também pode configurar o parâmetro `LaunchTemplateSpecification` na estrutura de substituições para atribuir um modelo de execução diferente a um conjunto de requisitos de instância, conforme necessário. Essa opção não está disponível atualmente no console. Para obter mais informações, consulte a LaunchTemplateOverridesReferência da API Amazon EC2 Auto Scaling.

Por padrão, você definiu o número de instâncias como da capacidade desejada do seu grupo do Auto Scaling.

Como alternativa, você pode definir o valor da capacidade desejada como o número de v CPUs ou a quantidade de memória. Para fazer isso, use a propriedade `DesiredCapacityType` na operação da API `CreateAutoScalingGroup` ou o campo suspenso Tipo de capacidade desejada no Console de gerenciamento da AWS. Essa é uma alternativa útil aos pesos de instância.

Flexibilidade de tipo da instância

Para aumentar a disponibilidade, implemente sua aplicação em vários tipos de instância. É uma prática recomendada usar vários tipos de instância para atender aos requisitos de capacidade. Dessa forma, o Amazon EC2 Auto Scaling pode iniciar outro tipo de instância se houver capacidade de instância insuficiente nas zonas de disponibilidade escolhidas.

Se houver capacidade de instância insuficiente com instâncias spot, o Amazon EC2 Auto Scaling continuará tentando iniciar a partir de outros pools de instâncias spot. (Os pools usados são determinados por sua escolha de tipos de instância e estratégia de alocação.) O Amazon EC2 Auto Scaling ajuda você a aproveitar a economia de custo das instâncias spot ao iniciá-las em vez de instâncias sob demanda.

Recomendamos ser flexível para pelo menos 10 tipos de instância para cada workload. Ao escolher seus tipos de instância, não se limite aos novos tipos de instância mais usados. Escolher tipos de instância de gerações mais antigas tende a resultar em menos interrupções spot, pois há menos demanda de clientes sob demanda.

Flexibilidade da zona de disponibilidade

Recomendamos fortemente que estenda seu grupo do Auto Scaling em várias zonas de disponibilidade. Com várias zonas de disponibilidade, você pode criar aplicações que executam o failover automaticamente entre as zonas para obter maior resiliência.

Como benefício adicional, você pode acessar um pool de capacidade mais profundo do Amazon EC2 em comparação com grupos em uma única zona de disponibilidade. Como a capacidade oscila independentemente para cada tipo de instância na zona de disponibilidade, é frequentemente possível obter maior capacidade computacional quando você tem tanto a flexibilidade de tipo de instância quanto da zona de disponibilidade.

Para ter mais informações sobre como usar várias Zonas de disponibilidade, consulte Exemplo: distribuir instâncias entre zonas de disponibilidade.

Preço máximo do spot

Ao criar seu grupo de Auto Scaling usando o AWS CLI ou um SDK, você pode especificar o parâmetro. `SpotMaxPrice` O parâmetro `SpotMaxPrice` determina o preço máximo que você está disposto a pagar por uma hora de instância spot.

Quando você especifica o parâmetro `WeightedCapacity` em suas substituições (ou `"DesiredCapacityType": "vcpu"` ou `"DesiredCapacityType": "memory-mib"` no nível do grupo), o preço máximo representa o preço unitário máximo, não o preço máximo de uma instância inteira.

É altamente recomendável que você não especifique um preço máximo. Talvez sua aplicação não seja executada se você não receber suas instâncias spot, como quando o preço máximo é muito baixo. Se você não especificar um preço máximo, o padrão será o preço sob demanda. Você pagará apenas o preço spot pelas instâncias spot que iniciar. Você ainda recebe os grandes descontos oferecidos pelas Instâncias spot. Esses descontos são possíveis devido ao preço spot estável disponível com o modelo de preço spot. Para obter mais informações, consulte Preços e economia no Manual do usuário do Amazon EC2 para instâncias do Linux.

Rebalanceamento proativo de capacidade

Se o seu caso de uso permitir, recomendamos o rebalanceamento de capacidade. O rebalanceamento de capacidade ajuda você a manter a disponibilidade da workload ao substituir proativamente as instâncias spot com risco de interrupção.

Quando a capacidade de rebalanceamento está habilitada, o Amazon EC2 Auto Scaling tenta substituir proativamente as instâncias spot que receberam uma recomendação de rebalanceamento de instância do EC2. É possível decidir rebalancear sua workload em instâncias spot novas ou existentes que não tenham um risco elevado de interrupção.

Para obter mais informações, consulte Rebalanceamento de capacidade no Auto Scaling para substituir instâncias spot em risco.

Comportamento do ajuste de escala

Quando você cria um grupo misto de instâncias, ele usa instâncias sob demanda por padrão. Para usar as instâncias spot, é necessário modificar a porcentagem do grupo a ser iniciada como instâncias sob demanda. Você pode especificar qualquer número de 0 a 100 para a porcentagem sob demanda.

Opcionalmente, você também pode designar um número base de instâncias sob demanda para começar. Se você fizer isso, o Amazon EC2 Auto Scaling aguardará para iniciar instâncias spot até depois de iniciar a capacidade básica de instâncias sob demanda quando o grupo for aumentado na escala horizontalmente. Depois de ultrapassada a capacidade básica, é usada a porcentagem sob demanda para determinar o número de instâncias spot e sob demanda que serão executadas.

O Amazon EC2 Auto Scaling converte o percentual para o número equivalente de instâncias. Se o resultado criar um número fracionário, ele arredonda para o próximo inteiro em favor das instâncias sob demanda.

A tabela a seguir demonstra o comportamento do grupo do Auto Scaling à medida que aumenta e diminui de tamanho.

Exemplo: comportamento de escalabilidade

Opções de compra	Tamanho de grupo e número total de instâncias em execução nas opções de compra			
	10	20	30	40
Exemplo 1: base de 10, 50/50% sob demanda/spot				

Opções de compra	Tamanho de grupo e número total de instâncias em execução nas opções de compra			
Instâncias sob demanda (valor base)	10	10	10	10
Instâncias sob demanda	0	5	10	15
Instâncias spot	0	5	10	15
Exemplo 2: base de 0, 0/100% sob demanda/spot				
Instâncias sob demanda (valor base)	0	0	0	0
Instâncias sob demanda	0	0	0	0
Instâncias spot	10	20	30	40
Exemplo 3: base de 0, 60/40% sob demanda/spot				
Instâncias sob demanda (valor base)	0	0	0	0
Instâncias sob demanda	6	12	18	24
Instâncias spot	4	8	12	16

Opções de compra	Tamanho de grupo e número total de instâncias em execução nas opções de compra			
Exemplo 4: base de 0, 100/0% sob demanda/spot				
Instâncias sob demanda (valor base)	0	0	0	0
Instâncias sob demanda	10	20	30	40
Instâncias spot	0	0	0	0
Exemplo 5: base de 12, 0/100% sob demanda/spot				
Instâncias sob demanda (valor base)	10	12	12	12
Instâncias sob demanda	0	0	0	0
Instâncias spot	0	8	18	28

Quando o tamanho do grupo aumenta, o Amazon EC2 Auto Scaling tenta equilibrar sua capacidade uniformemente em suas zonas de disponibilidade especificadas. Em seguida, ele inicia os tipos de instância de acordo com a estratégia de alocação especificada.

Quando o tamanho do grupo diminui, o Amazon EC2 Auto Scaling primeiro identifica qual dos dois tipos (spot ou sob demanda) deve ser encerrado. Em seguida, ele tenta encerrar as instâncias de forma balanceada nas zonas de disponibilidade especificadas. Também favorece o encerramento de instâncias de uma forma que se alinhe mais às suas estratégias de alocação. Para obter mais

informações sobre políticas de encerramento, consulte Trabalhar com políticas de término do Amazon EC2 Auto Scaling.

Disponibilidade regional dos tipos de instância

A disponibilidade dos tipos de instância do EC2 varia de acordo com sua Região da AWS. Por exemplo, os tipos de instância de geração mais recente podem ainda não estar disponíveis em uma determinada região. Devido às variações na disponibilidade de instâncias entre regiões, você pode encontrar problemas ao fazer solicitações programáticas se vários tipos de instância em suas substituições não estiverem disponíveis em sua região. Usar vários tipos de instância que não estão disponíveis na sua região pode fazer com que a solicitação falhe completamente. Para resolver o problema, repita a solicitação com diferentes tipos de instância, certificando-se de que cada tipo de instância esteja disponível na região. Para pesquisar os tipos de instância oferecidos por localização, use o describe-instance-type-offeringscomando. Para obter mais informações, consulte Tipos de nomes do host de instâncias do Amazon EC2 no Guia do usuário do Amazon EC2.

Recursos relacionados

Para obter as práticas recomendadas para Instâncias spot, consulte Práticas recomendadas para o EC2 Spot no Guia do usuário do Amazon EC2 para instâncias do Linux.

Limitações

Depois de adicionar substituições a um grupo do Auto Scaling usando uma política de instâncias mistas, você pode atualizar as substituições com a `UpdateAutoScalingGroup`chamada de API, mas não excluí-las. Para remover completamente as substituições, é necessário primeiro mudar o grupo do Auto Scaling para usar um modelo de execução ou uma configuração de execução em vez de uma política de instâncias mistas. Em seguida, você pode adicionar uma política de instâncias mistas novamente sem nenhuma substituição.

Estratégias de alocação para vários tipos de instância

Ao usar vários tipos de instância, você gerencia como o Amazon EC2 Auto Scaling atenderá à sua capacidade sob demanda e spot dos tipos de instância possíveis. Para fazer isso, você especifica estratégias de alocação.

Para analisar as etapas de configuração de um grupo de instâncias mistas, consulte Visão geral da configuração para criar um grupo de instâncias mistas.

Conteúdo

- Instâncias spot

- Instâncias sob demanda
- Como as estratégias de alocação funcionam com ponderações

Instâncias spot

O Amazon EC2 Auto Scaling fornece as seguintes estratégias de alocação para instâncias spot:

`price-capacity-optimized` (recomendado)

A estratégia de alocação otimizada de preço e capacidade analisa o preço e a capacidade para selecionar os pools de instâncias spot com menor probabilidade de interrupção e com o preço mais baixo possível.

Recomendamos esta estratégia quando você está começando. Para obter mais informações, consulte Introdução à estratégia de price-capacity-optimized alocação para instâncias spot do EC2 no AWS blog.

`capacity-optimized`

O Amazon EC2 Auto Scaling solicita sua instância spot do pool com capacidade ideal para o número de instâncias que estão sendo executadas.

Com as instâncias spot, a definição de preço muda lentamente ao longo do tempo com base em tendências de longo prazo na oferta e na demanda. No entanto, a capacidade flutua em tempo real. A estratégia `capacity-optimized` executa Instâncias spot automaticamente nos grupos mais disponíveis observando dados de capacidade em tempo real e prevendo quais são os mais disponíveis. Isso ajuda a minimizar possíveis interrupções para cargas de trabalho que podem ter um custo mais alto de interrupção associado ao reinício do trabalho e ao ponto de verificação. Para dar a certos tipos de instância uma maior chance de serem executadas primeiro, use `capacity-optimized-prioritized`.

`capacity-optimized-prioritized`

Você define a ordem dos tipos de instância para as substituições do modelo de execução da prioridade mais alta para a mais baixa (do primeiro ao último na lista). O Amazon EC2 Auto Scaling honra as prioridades de tipo de instância com base no melhor esforço, mas otimiza primeiro a capacidade. Essa é uma boa opção para workloads em que a possibilidade de interrupção deve ser minimizada, mas em que a preferência por determinados tipos de instância também é importante. Se a estratégia de alocação sob demanda for definida como `prioritized`, a mesma prioridade será aplicada ao atender a capacidade sob demanda.

`lowest-price` (não recomendado)

> **Warning**
>
> Não recomendamos a estratégia de `lowest-price` porque ela representa o maior risco de interrupção para as instâncias spot.

O Amazon EC2 Auto Scaling solicita suas instâncias spot usando os pools de menor preço dentro de uma zona de disponibilidade, entre o número N de pools spot que você especifica para a configuração de pools de menor preço. Por exemplo, se você especificar quatro tipos de instância e quatro zonas de disponibilidade, seu grupo do Auto Scaling poderá acessar até 16 pools spot. (Quatro em cada zona de disponibilidade.) Se você especificar dois pools de spot (N=2) para a estratégia de alocação, seu grupo do Auto Scaling poderá aproveitar os dois pools de preço mais baixo por zona de disponibilidade para preencher sua capacidade spot.

A estratégia de `lowest-price` só está disponível com o uso da AWS CLI.

O Amazon EC2 Auto Scaling tenta extrair instâncias spot do número N de pools que você especifica. No entanto, se um pool ficar sem capacidade spot antes de atender à capacidade desejada, o Amazon EC2 Auto Scaling continua a atender à sua solicitação usando o próximo pool de preço mais baixo. Para atender à capacidade desejada, você pode receber instâncias spot de mais pools do que o número N especificado. Da mesma forma, se a maioria dos pools não tiver capacidade spot, você poderá receber a capacidade total desejada de menos pools do que o número N especificado.

> **Note**
>
> Se você configurar sua instância spot para iniciar com AMD SEV-SNP ativado, uma tarifa adicional de uso por hora será cobrada. Essa tarifa equivale a 10% da Taxa sob demanda por hora do tipo de instância selecionado. Se a estratégia de alocação usar o preço como entrada, a Amazon EC2 Auto Scaling não incluirá essa tarifa adicional; somente o preço spot será usado.

Instâncias sob demanda

O Amazon EC2 Auto Scaling fornece as seguintes estratégias de alocação que podem ser usadas para instâncias sob-demanda:

`lowest-price`

O Amazon EC2 Auto Scaling implanta automaticamente o tipo de instância com preço mais baixo em cada zona de disponibilidade com base no preço sob demanda atual.

Para atender à capacidade desejada, você pode receber instâncias sob demanda de mais de um tipo de instância em cada zona de disponibilidade. Isso depende da quantidade de capacidade que você solicitar.

`prioritized`

Ao atender à capacidade sob demanda, o Amazon EC2 Auto Scaling determina qual tipo de instância usar primeiro com base na ordem dos tipos de instância na lista de substituições de modelo de execução. Por exemplo, digamos que você especifique três substituições de modelo de execução na seguinte ordem: `c5.large`, `c4.large` e `c3.large`. Quando suas instâncias sob demanda são iniciadas, o grupo do Auto Scaling preenche a capacidade sob demanda começando com,`c5.large` `c4.large` e, em seguida, `c3.large`.

Considere o seguinte ao gerenciar a ordem de prioridade de suas instâncias sob demanda:

- Você pode pagar antecipadamente pelo uso para obter descontos significativos para Instâncias sob demanda usando Savings Plans ou instâncias reservadas. Para obter mais informações, consulte a página de preços do Amazon EC2.
- Com instâncias reservadas, sua taxa de desconto da definição de preço normal da instância sob demanda se aplicará se o Amazon EC2 Auto Scaling iniciar tipos de instância correspondentes. Portanto, se você tiver Instâncias reservadas não utilizadas para `c4.large`, poderá definir a prioridade do tipo de instância para dar a prioridade mais alta para suas Instâncias reservadas a um tipo de instância `c4.large`. Quando uma instância `c4.large` é ativada, você recebe os preços de instância reservada.
- Com os Savings Plans, sua taxa de desconto da definição de preço normal da instância sob demanda é aplicada ao usar os Amazon EC2 Instance Savings Plans ou Compute Savings Plans. Com Savings Plans, você tem mais flexibilidade ao priorizar seus tipos de instância. Contanto que você use tipos de instância cobertos pelo seu Savings Plans, é possível defini-los em qualquer ordem de prioridade. Você também pode ocasionalmente alterar toda a ordem

de seus tipos de instância, enquanto ainda recebe a taxa de desconto de Savings Plans. Para obter mais informações sobre Savings Plans, consulte o Savings Plans User Guide (Guia do usuário de Savings Plans).

Como as estratégias de alocação funcionam com ponderações

Quando você especifica o parâmetro `WeightedCapacity` em suas substituições (ou `"DesiredCapacityType": "vcpu"` ou `"DesiredCapacityType": "memory-mib"` no nível do grupo), as estratégias de alocação funcionam exatamente como em outros grupos do Auto Scaling.

Suponha que você tenha um grupo de Auto Scaling com vários tipos de instância que têm quantidades variáveis de v. CPUs Você usa `lowest-price` para suas estratégias de alocação spot e sob demanda. Se você optar por atribuir pesos com base na contagem de vCPUs de cada tipo de instância, o Amazon EC2 Auto Scaling iniciará os tipos de instância que tenham o menor preço por valores de peso atribuídos (por exemplo, por vCPU) no momento do cumprimento. Se for uma instância spot, isso significa o menor preço spot por vCPU. Se for uma instância sob demanda, isso significa o menor preço sob demanda por vCPU.

Para obter mais informações, consulte Configurar um grupo do Auto Scaling para usar ponderação da instância.

Crie um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos

Em vez de escolher manualmente os tipos de instância para seu grupo de instâncias mistas, você pode especificar um conjunto de atributos de instância que descrevem seus requisitos de computação. À medida que o Amazon EC2 Auto Scaling inicia as instâncias, todos os tipos de instância usados pelo grupo do Auto Scaling devem corresponder aos atributos de instância exigidos. Isso é conhecido como seleção de tipo de instância baseada em atributos.

Essa abordagem é ideal para workloads e frameworks que podem ser flexíveis sobre quais tipos de instância são usadas, como contêineres, big data e CI/CD.

Os benefícios da seleção de tipo de instância baseada em atributos são os seguintes:

- Flexibilidade ideal para instâncias spot: O Amazon EC2 Auto Scaling pode selecionar entre uma ampla variedade de tipos de instância para iniciar instâncias spot. Isso atende à prática recomendada do spot de ser flexível em relação aos tipos de instância, o que dá ao serviço

Spot do Amazon EC2 uma chance melhor de encontrar e alocar a quantidade necessária de capacidade computacional.

- Use facilmente os tipos de instâncias certos: com tantos tipos de instância disponíveis, encontrar os tipos de instância corretos para a workload pode ser demorado. Se você especificar os atributos de instância, os tipos de instância terão automaticamente os atributos necessários para sua workload.
- Uso automático de novos tipos de instância: Seus grupos do Auto Scaling podem usar tipos de instância de geração mais nova, à medida que são executadas. Tipos de instância de geração mais nova são usados automaticamente quando correspondem aos seus requisitos e se alinham com as estratégias de alocação escolhidas para o grupo do Auto Scaling.

Tópicos

- Como funciona a seleção de tipo de instância baseada em atributos
- Proteção de preço
- Proteção de performance
- Pré-requisitos
- Crie um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos
- Crie um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos (AWS CLI)
- Exemplo de configuração
- Pré-visualize os tipos de instância
- Recursos relacionados

Como funciona a seleção de tipo de instância baseada em atributos

Com a seleção de tipo de instância baseada em atributos, em vez de fornecer uma lista de tipos de instância específicos, você fornece uma lista de atributos de instância que suas instâncias exigem, como:

- Contagem de vCPU — O número mínimo e máximo de v CPUs por instância.
- Memória — O mínimo e o máximo GiBs de memória por instância.
- Armazenamento local: se o sistema deve usar o EBS ou volumes de armazenamento de instâncias para armazenamento local.

- Performance intermitente: se o sistema deve usar a família de instâncias T, incluindo os tipos T4g, T3a, T3 e T2.

Há muitas opções disponíveis para definir os requisitos de sua instância. Para obter uma descrição de cada opção e os valores padrão, consulte a Referência da InstanceRequirementsAPI do Amazon EC2 Auto Scaling.

Quando seu grupo do Auto Scaling precisar iniciar uma instância, ele procurará por tipos de instância que correspondam aos atributos especificados e que estejam disponíveis nessa zona de disponibilidade. As estratégias de alocação determinam quais tipos de instância serão executados. Por padrão, a seleção de tipo de instância baseada em atributos tem um atributo de proteção de preço ativado para impedir que o grupo do Auto Scaling inicie tipos de instância que excedam seus limites de orçamento.

Por padrão, você usa o número de instâncias como unidade de medida ao definir a capacidade desejada do seu grupo de Auto Scaling, o que significa que cada instância conta como uma unidade.

Como alternativa, você pode definir o valor da capacidade desejada como o número de v CPUs ou a quantidade de memória. Para fazer isso, use o campo suspenso Tipo de capacidade desejada na `DesiredCapacityType` propriedade Console de gerenciamento da AWS ou na operação da `CreateAutoScalingGroup` `UpdateAutoScalingGroup` API. Em seguida, o Amazon EC2 Auto Scaling executa o número de instâncias necessárias para atender à capacidade desejada de vCPU ou memória. Por exemplo, se você usar v CPUs como o tipo de capacidade desejado e usar instâncias com 2 v CPUs cada, uma capacidade desejada de 10 v CPUs iniciaria 5 instâncias. Essa é uma alternativa útil aos pesos de instância.

Proteção de preço

Com a proteção de preço, você pode especificar o preço máximo que está disposto a pagar pelas instâncias do EC2 executadas pelo seu grupo do Auto Scaling. A proteção de preços é um recurso que impede que sua frota spot use tipos de instância que você consideraria muito caros, mesmo que atendam aos atributos especificados.

A proteção de preço é ativada por padrão e tem limites de preço separados para instâncias sob Demanda e instâncias spot. Quando o Amazon EC2 Auto Scaling precisa iniciar novas instâncias, os tipos de instância com preço acima do limite relevante não são iniciados.

Tópicos

- Proteção de preço de instância sob demanda

- Proteção de preço spot
- Personalizar a proteção de preços

Proteção de preço de instância sob demanda

Para instâncias sob demanda, você define o preço máximo sob demanda que está disposto a pagar como uma porcentagem maior do que um preço sob demanda identificado. O preço sob demanda identificado é o preço do tipo de instância C, M ou R da geração atual com o menor preço e com os atributos especificados.

Se um valor de proteção de preço sob demanda não for explicitamente definido, será usado um preço sob demanda máximo padrão 20% maior do que o preço sob demanda identificado.

Proteção de preço spot

Por padrão, o Amazon EC2 Auto Scaling aplicará automaticamente a proteção de preço de instância spot ideal para selecionar de forma consistente entre uma ampla variedade de tipos de instância. Você também pode definir manualmente a proteção de preço. No entanto, deixar que o Amazon EC2 Auto Scaling faça isso por você pode aumentar a probabilidade de que sua capacidade de spot seja atendida.

É possível especificar manualmente a proteção de preço usando uma das opções a seguir. Se você definir manualmente a proteção de preço, recomendamos usar a primeira opção.

- Uma porcentagem de um preço sob demanda identificado: O preço sob demanda identificado é o preço do tipo de instância C, M ou R da geração atual com o menor preço e com os atributos especificados.
- Uma porcentagem maior que um preço spot identificado: O preço spot identificado é o preço do tipo de instância C, M ou R da geração atual com o menor preço e com os atributos especificados. Não é recomendável usar esse parâmetro, pois os preço spot podem flutuar e, portanto, seu limite de proteção de preço também poderá flutuar.

Personalizar a proteção de preços

Você pode personalizar os limites de proteção de preços no console Amazon EC2 Auto Scaling ou usando o ou. AWS CLI SDKs

- No console, use as configurações de Proteção de preço sob demanda e Proteção de preço spot em Atributos adicionais da instância.

- Na InstanceRequirementsestrutura, para especificar o limite de proteção de preço da instância sob demanda, use a `OnDemandMaxPricePercentageOverLowestPrice` propriedade. Para especificar o limite de proteção de preço da instância spot, use a propriedade `MaxSpotPriceAsPercentageOfOptimalOnDemandPrice` ou `SpotMaxPricePercentageOverLowestPrice`.

Se você definir o tipo de capacidade desejado (`DesiredCapacityType`) como v CPUs ou GiB de memória, a proteção de preço será aplicada com base no preço por vCPU ou por memória, em vez do preço por instância.

Você também pode desativar a proteção de preços. Para indicar que não há limite de proteção de preço, especifique um valor percentual alto, como 999999.

> **Note**
>
> Se nenhum tipo de instância C, M ou R da geração atual corresponder aos atributos especificados, a proteção de preço ainda será aplicável. Quando nenhuma correspondência é encontrada, o preço identificado é o dos tipos de instância da geração atual com o menor preço ou, se não houver, dos tipos de instância da geração anterior com o menor preço que correspondam a seus atributos.

Proteção de performance

A Proteção de performance é um recurso para garantir que seu grupo do Auto Scaling use tipos de instância que sejam semelhantes ou excedam uma linha de base de performance especificada. Para usar a proteção de performance, especifique uma família de instâncias como referência de linha de base. Os recursos da família de instâncias especificada estabelecem o nível de performance mais baixo aceitável. Quando o Auto Scaling seleciona tipos de instância, ele considera seus atributos e a linha de base de performance especificados. Os tipos de instância que estão abaixo da linha de base de performance são automaticamente excluídos da seleção, mesmo que correspondam aos outros atributos especificados. Isso garante que todos os tipos de instância selecionados ofereçam performance semelhante ou melhor do que a linha de base estabelecida pela família de instâncias especificada. O Auto Scaling usa essa linha de base para orientar a seleção do tipo de instância, mas não há garantia de que os tipos de instância selecionados sempre excederão a linha de base de cada aplicação.

No momento, esse recurso oferece suporte apenas à performance da CPU como fator de performance da linha de base. A performance da CPU da família de instâncias especificada serve como linha de base de performance, garantindo que os tipos de instância selecionados sejam semelhantes ou excedam essa linha de base. Famílias de instâncias com os mesmos processadores de CPU levam aos mesmos resultados de filtragem, mesmo que a performance da rede ou do disco seja diferente. Por exemplo, especificar `c6in` ou `c6i` como referência de linha de base produziria resultados de filtragem idênticos com base na performance porque as duas famílias de instâncias usam o mesmo processador de CPU.

Famílias de instâncias não compatíveis

As seguintes famílias de instâncias não são compatíveis com a proteção de performance:

- `c1`
- `g3` | `g3s`
- `hpc7g`
- `m1` | `m2`
- `mac1` | `mac2` | `mac2-m1ultra` | `mac2-m2` | `mac2-m2pro`
- `p3dn` | `p4d` | `p5`
- `t1`
- `u-12tb1` | `u-18tb1` | `u-24tb1` | `u-3tb1` | `u-6tb1` | `u-9tb1` | `u7i-12tb` | `u7in-16tb` | `u7in-24tb` | `u7in-32tb`

Se você habilitar a proteção de performance especificando uma família de instâncias compatível, os tipos de instância retornados excluirão as famílias de instâncias incompatíveis acima.

Exemplo: definir uma linha de base de performance da CPU

No exemplo a seguir, o requisito da instância é ser iniciada com tipos de instância que tenham núcleos de CPU com a mesma performance da família de instâncias `c6i`. Isso filtrará os tipos de instância com processadores de CPU de menor desempenho, mesmo que atendam aos outros requisitos de instância especificados, como o número de v. CPUs Por exemplo, se os atributos de instância especificados incluírem 4 v CPUs e 16 GB de memória, um tipo de instância com esses atributos, mas com desempenho de CPU inferior ao que `c6i` será excluído da seleção.

```
"BaselinePerformanceFactors": {
        "Cpu": {
```

```
            "References": [
                {
                    "InstanceFamily": "c6i"
                }
            ]
        }
```

Considerações

Considere o seguinte ao usar a proteção de performance:

- É possível especificar tipos de instância ou atributos de instância, mas não os dois ao mesmo tempo.
- É possível especificar, no máximo, quatro estruturas de `InstanceRequirements` em uma configuração de solicitação.

Pré-requisitos

- Criar um modelo de execução. Para obter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.
- Verifique se o modelo de execução já não solicita instâncias spot.

Crie um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos

Use o procedimento a seguir para criar um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos. Para ajudar a percorrer as etapas com eficiência, algumas seções opcionais são ignoradas.

Para a maioria das cargas de trabalho de uso geral, basta especificar o número de v CPUs e a memória de que você precisa. Para casos de uso avançados, você pode especificar atributos como tipo de armazenamento, interfaces de rede, fabricante da CPU e tipo de acelerador.

Para analisar as etapas de configuração de um grupo de instâncias mistas, consulte Visão geral da configuração para criar um grupo de instâncias mistas.

Para criar um grupo de instâncias mistas

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.

2. Na barra de navegação na parte superior da tela, selecione a mesma Região da AWS usada na criação do modelo de execução.
3. Selecione Criar um grupo do Auto Scaling.
4. Na página Choose launch template or configuration (Escolher o modelo ou a configuração de execução), em Auto Scaling group name (Nome do grupo do Auto Scaling) insira um nome para o grupo do Auto Scaling.
5. Para escolher o modelo de inicialização, faça o seguinte:

   a. Em Launch template (Modelo de execução), escolha um modelo de execução existente.

   b. Em Launch template version (Versão do modelo de execução), indique se o grupo do Auto Scaling usará a versão padrão, a mais recente ou uma versão específica do modelo de execução no aumento da escala na horizontal.

   c. Verifique se o modelo de execução oferece suporte a todas as opções que você está planejando usar e escolha Next (Próximo).
6. Na página Escolher opções de inicialização da instância, faça o seguinte:

   a. Para Requisitos de tipo de instância, selecione Substituir modelo de execução.

   > **Note**
   >
   > Se você escolher um modelo de execução que já contenha um conjunto de atributos de instância, como vCPUs e memória, os atributos da instância serão exibidos. Esses atributos são adicionados às propriedades do grupo do Auto Scaling, onde você pode atualizá-los do console no Amazon EC2 Auto Scaling, a qualquer momento.

   b. Em Especificar atributos da instância, comece inserindo seus requisitos de v CPUs e memória.

      - Para v CPUs, insira o número mínimo e máximo desejado de CPUs v. Para especificar nenhum limite, selecione Sem mínimo, Sem máximo ou ambos.
      - Em Memory (GiB) (Memória), insira a quantidade mínima e máxima de memória desejada. Para não especificar nenhum limite, selecione No minimum (Sem mínimo), No maximum (Sem máximo) ou ambos.

c. (Opcional) Em Additional instance attributes (Atributos de instância adicionais), você pode, opcionalmente, especificar um ou mais atributos para expressar seus requisitos de computação com mais detalhes. Cada atributo adicional inclui mais restrições à solicitação.

d. (Opcional) Expanda Preview matching instance types (Previsualizar os tipos de instância correspondentes) para visualizar os tipos de instância que têm os atributos especificados.

e. Em Opções de compra de instâncias, para Distribuição de instâncias, especifique as porcentagens do grupo para lançamento como instâncias sob demanda e como instâncias spot. Se sua aplicação for sem estado, tolerante a falhas e puder lidar com a interrupção de uma instância, você poderá especificar uma porcentagem maior de instâncias spot.

f. (Opcional) Quando você especifica uma porcentagem para instâncias spot, selecione Incluir capacidade básica sob demanda e depois especifique a capacidade inicial mínima do grupo do Auto Scaling que deve ser atendido por instâncias sob demanda. Se a capacidade básica for ultrapassada, as configurações Instances distribution (Distribuição de instâncias) serão usadas para determinar quantas instâncias spot e instâncias sob demanda serão executadas.

g. Sob Allocation strategies (Estratégias de alocação), Lowest price (Preço mais baixo) é selecionado automaticamente para a On-Demand allocation strategy (Estratégia de alocação sob demanda), e não pode ser alterado.

h. Para Spot allocation strategy (Estratégia de alocação spot), selecione uma estratégia de alocação. A capacidade de preço otimizada é selecionada por padrão.

i. Em Rebalanceamento de capacidade, escolha se você deseja habilitar ou desabilitar o rebalanceamento de capacidade. Use o rebalanceamento de capacidade para responder automaticamente quando suas instâncias spot se aproximarem do encerramento de uma interrupção spot. Para obter mais informações, consulte Rebalanceamento de capacidade no Auto Scaling para substituir instâncias spot em risco.

j. Em Network (Rede), para VPC, escolha uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução.

k. Para Availability Zones and subnets (Zonas de disponibilidade e sub-redes), selecione uma ou mais sub-redes na VPC especificada. Use sub-redes em várias zonas de disponibilidade para alta disponibilidade. Para obter mais informações, consulte Considerações sobre a escolha de sub-redes da VPC.

l. Escolha Avançar, Avançar.

7. Na etapa Configurar políticas de escalabilidade e tamanho do grupo, faça o seguinte:

a. Para medir a capacidade desejada em unidades que não sejam instâncias, escolha a opção apropriada para Tamanho do grupo, Tipo de capacidade desejada. UnidadesCPUs, v e Memória GiB são suportadas. Por padrão, o Amazon EC2 Auto Scaling especifica Unidades, o que quer dizer número de instâncias.

b. Para Capacidade desejada, o tamanho inicial do seu grupo do Auto Scaling.

c. Na seção Escalabilidade, em Limites de escalabilidade, se o novo valor para Capacidade desejada for maior que Capacidade mínima desejada e Capacidade máxima desejada, a Capacidade máxima desejada será automaticamente aumentada para o novo valor da capacidade desejada. É possível alterar esses limites conforme necessário. Para obter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.

8. Escolha Skip to review (Ir para revisão).
9. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Crie um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos (AWS CLI)

Para criar um grupo de instâncias mistas usando a linha de comando

Use um dos seguintes comandos:

- create-auto-scaling-group (AWS CLI)
- Novo- ASAuto ScalingGroup (AWS Tools for Windows PowerShell)

Exemplo de configuração

Para criar um grupo de Auto Scaling com seleção de tipo de instância baseada em atributos usando o AWS CLI, use o comando a seguir. create-auto-scaling-group

Os seguintes atributos de instância são especificados:

- `VCpuCount`— Os tipos de instância devem ter no mínimo quatro v CPUs e no máximo oito CPUs v.
- `MemoryMiB` – os tipos de instância devem ter no mínimo 16.384 MiB de memória.
- `CpuManufacturers` – os tipos de instância devem ter uma CPU fabricada pela Intel.

JSON

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Veja a seguir um exemplo de arquivo `config.json`.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredCapacityType": "units",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Default"
            },
            "Overrides": [{
                "InstanceRequirements": {
                    "VCpuCount": {"Min": 4, "Max": 8},
                    "MemoryMiB": {"Min": 16384},
                    "CpuManufacturers": ["intel"]
                }
            }]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 50,
            "SpotAllocationStrategy": "price-capacity-optimized"
        }
    },
    "MinSize": 0,
    "MaxSize": 100,
    "DesiredCapacity": 4,
    "DesiredCapacityType": "units",
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

Para definir o valor da capacidade desejada como o número de v CPUs ou a quantidade de memória, especifique `"DesiredCapacityType": "vcpu"` ou `"DesiredCapacityType": "memory-mib"` no arquivo. O tipo de capacidade padrão desejado é `units`, que define o valor da capacidade desejada como o número de instâncias.

YAML

Como alternativa, você pode usar o create-auto-scaling-groupcomando a seguir para criar o grupo Auto Scaling. Isso faz referência a um arquivo YAML como o único parâmetro para seu grupo do Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Veja a seguir um exemplo de arquivo `config.yaml`.

```
---
AutoScalingGroupName: my-asg
DesiredCapacityType: units
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceRequirements:
        VCpuCount:
          Min: 2
          Max: 4
        MemoryMiB:
          Min: 2048
        CpuManufacturers:
        - intel
  InstancesDistribution:
    OnDemandPercentageAboveBaseCapacity: 50
    SpotAllocationStrategy: price-capacity-optimized
MinSize: 0
MaxSize: 100
DesiredCapacity: 4
DesiredCapacityType: units
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Para definir o valor da capacidade desejada como o número de v CPUs ou a quantidade de memória, especifique `DesiredCapacityType: vcpu` ou `DesiredCapacityType: memory-mib` no arquivo. O tipo de capacidade padrão desejado é `units`, que define o valor da capacidade desejada como o número de instâncias.

Para ver um exemplo de como usar vários modelos de execução com seleção de tipo de instância baseada em atributos, consulte Usar vários modelos de execução.

Pré-visualize os tipos de instância

É possível previsualizar os tipos de instância que correspondem aos requisitos de computação sem iniciá-los e ajustar seus requisitos, se necessário. Ao criar o grupo do Auto Scaling no console do Amazon EC2 Auto Scaling, uma previsualização dos tipos de instância aparece na seção Preview matching instance types (Previsualize os tipos de instância correspondentes) na página Choose instance launch options (Escolha as opções de execução da instância).

Como alternativa, você pode visualizar os tipos de instância fazendo uma chamada de GetInstanceTypesFromInstanceRequirementsAPI do Amazon EC2 usando o AWS CLI ou um SDK. Transmita os parâmetros `InstanceRequirements` na solicitação, no formato exato que você usaria para criar ou atualizar um grupo do Auto Scaling. Para mais informações, consulte Preview instance types with specified attributes (Previsualize tipos de instância com atributos especificados) no Amazon EC2 User Guide (Guia do usuário do Amazon EC2).

Recursos relacionados

Para saber mais sobre a seleção de tipo de instância baseada em atributos, consulte Seleção de tipo de instância baseada em atributos para EC2 Auto Scaling e EC2 Fleet no blog. AWS

Você pode declarar a seleção de tipo de instância baseada em atributos ao criar um grupo do Auto Scaling usando CloudFormation. Para obter mais informações, consulte o trecho de exemplo na seção Trechos de modelo de escalonamento automático do CloudFormation Guia do usuário.

Criar um grupo misto de instâncias escolhendo manualmente os tipos de instância

Este tópico mostra como iniciar vários tipos de instância em um único grupo do Auto Scaling escolhendo manualmente seus tipos de instância.

Se você preferir usar atributos de instância como critérios para selecionar tipos de instância, consulte Crie um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos.

Conteúdos

- Pré-requisitos
- Criar um grupo de instâncias mistas (console)
- Criar um grupo de instâncias mistas (AWS CLI)

- Exemplos de configuração

Pré-requisitos

- Criar um modelo de execução. Para obter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.
- Verifique se o modelo de execução já não solicita instâncias spot.

Criar um grupo de instâncias mistas (console)

Use o procedimento a seguir para criar um grupo de instâncias mistas escolhendo manualmente quais tipos de instância seu grupo pode iniciar. Para ajudar a percorrer as etapas com eficiência, algumas seções opcionais são ignoradas.

Para analisar as etapas de configuração de um grupo de instâncias mistas, consulte Visão geral da configuração para criar um grupo de instâncias mistas.

Para criar um grupo de instâncias mistas

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, selecione a mesma Região da AWS usada na criação do modelo de execução.
3. Selecione Criar um grupo do Auto Scaling.
4. Na página Choose launch template or configuration (Escolher o modelo ou a configuração de execução), em Auto Scaling group name (Nome do grupo do Auto Scaling) insira um nome para o grupo do Auto Scaling.
5. Para escolher o modelo de inicialização, faça o seguinte:
    a. Em Launch template (Modelo de execução), escolha um modelo de execução existente.
    b. Em Launch template version (Versão do modelo de execução), indique se o grupo do Auto Scaling usará a versão padrão, a mais recente ou uma versão específica do modelo de execução no aumento da escala na horizontal.
    c. Verifique se seu modelo de execução oferece suporte a todas as opções que você planeja usar e, em seguida, escolha Next (Próximo).
6. Na página Escolha as opções de iniciar uma instância, faça o seguinte:

a. Para Instance type requirements (Requisitos de tipo de instância), selecione Override launch template (Substituir modelo de execução), e depois escolha Manually add instance types (Adicionar tipos de instância manualmente).

b. Escolha os tipos de instância. Você pode usar nossas recomendações como ponto de partida. A opção de família e geração flexível é selecionada por padrão.

   - Para alterar a ordem dos tipos de instância, use as setas. Se você escolher uma estratégia de alocação compatível com priorização, a ordem do tipo de instância definirá sua prioridade de execução.
   - Para remover um tipo de instância, escolha X.
   - (Opcional) Para as caixas na coluna Peso, atribua um peso relativo a cada tipo de instância. Para fazer isso, insira o número de unidades que uma instância desse tipo conta para a capacidade desejada do grupo. Isso pode ser útil se os tipos de instância oferecerem diferentes recursos de vCPU, memória, armazenamento ou largura de banda de rede. Para obter mais informações, consulte Configurar um grupo do Auto Scaling para usar ponderação da instância.

     Se você optar por usar as recomendações flexíveis de tamanho, todos os tipos de instância que fazem parte desta seção terão automaticamente um valor de peso. Se você não quiser especificar nenhum peso, desmarque as caixas na coluna Peso para todos os tipos de instância.

c. Em Instance purchase options (Opções de compra), para Instances distribution (Distribuição de instâncias), especifique as porcentagens de instâncias do grupo a serem iniciadas como instâncias sob demanda e instâncias spot, respectivamente. Se a aplicação for sem estado, tolerante a falhas e puder lidar com uma interrupção de instância, você poderá especificar uma porcentagem maior de instâncias spot.

d. (Opcional) Quando você especifica uma porcentagem para instâncias spot, selecione Incluir capacidade básica sob demanda e depois especifique a capacidade inicial mínima do grupo do Auto Scaling que deve ser atendido por instâncias sob demanda. Se a capacidade básica for ultrapassada, as configurações Instances distribution (Distribuição de instâncias) serão usadas para determinar quantas instâncias spot e instâncias sob demanda serão executadas.

e. Em Allocation strategies (Estratégias de alocação), para On-Demand allocation strategy (Estratégia de alocação sob demanda), selecione uma estratégia de alocação. Quando

você escolhe manualmente seus tipos de instância, a opção Priorizada é selecionada por padrão.

f. Para Spot allocation strategy (Estratégia de alocação spot), selecione uma estratégia de alocação. A capacidade de preço otimizada é selecionada por padrão.

   Se você escolheu Capacidade otimizada, você pode, opcionalmente, marcar a caixa Priorizar tipos de instância para permitir que o Amazon EC2 Auto Scaling escolha qual tipo de instância iniciar primeiro com base na ordem em que seus tipos de instância estão listados.

g. Em Rebalanceamento de capacidade, escolha se você deseja habilitar ou desabilitar o rebalanceamento de capacidade. Use o rebalanceamento de capacidade para responder automaticamente quando suas instâncias spot se aproximarem do encerramento de uma interrupção spot. Para obter mais informações, consulte Rebalanceamento de capacidade no Auto Scaling para substituir instâncias spot em risco.

h. Em Network (Rede), para VPC, escolha uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução.

i. Para Availability Zones and subnets (Zonas de disponibilidade e sub-redes), selecione uma ou mais sub-redes na VPC especificada. Use sub-redes em várias zonas de disponibilidade para alta disponibilidade. Para obter mais informações, consulte Considerações sobre a escolha de sub-redes da VPC.

j. Escolha Avançar, Avançar.

7. Na etapa Configurar políticas de escalabilidade e tamanho do grupo, faça o seguinte:

   a. Em tamanho de grupo para Capacidade desejada, insira o número inicial de instâncias a serem executadas.

      Por padrão, a capacidade desejada é expressa como o número de instâncias. Se você atribuiu pesos aos seus tipos de instância, deverá converter esse valor na mesma unidade de medida usada para atribuir pesos, como o número de v. CPUs

   b. Na seção Escalabilidade, em Limites de escalabilidade, se o novo valor para Capacidade desejada for maior que Capacidade mínima desejada e Capacidade máxima desejada, a Capacidade máxima desejada será automaticamente aumentada para o novo valor da capacidade desejada. É possível alterar esses limites conforme necessário. Para obter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.

8. Escolha Skip to review (Ir para revisão).

9. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Criar um grupo de instâncias mistas (AWS CLI)

Para criar um grupo de instâncias mistas usando a linha de comando

Use um dos seguintes comandos:

- create-auto-scaling-group (AWS CLI)
- Novo- ASAuto ScalingGroup (AWS Tools for Windows PowerShell)

Exemplos de configuração

Os exemplos de configuração a seguir mostram como iniciar instâncias spot usando as diferentes estratégias de alocação spot.

> **Note**
>
> Esses exemplos mostram como usar um arquivo de configuração formatado em JSON ou YAML. Se você usar a AWS CLI versão 1, deverá especificar um arquivo de configuração formatado em JSON. Se você usar a AWS CLI versão 2, poderá especificar um arquivo de configuração formatado em YAML ou JSON.

Exemplos

- Exemplo 1: Iniciar instâncias spot usando a estratégia de alocação capacity-optimized
- Exemplo 2: Iniciar instâncias spot usando a estratégia de alocação capacity-optimized-prioritized
- Exemplo 3: Iniciar instâncias spot usando a estratégia de alocação lowest-price diversificada em dois grupos
- Exemplo 4: Iniciar Instâncias spot usando a estratégia de alocação price-capacity-optimized

Exemplo 1: Iniciar instâncias spot usando a estratégia de alocação **`capacity-optimized`**

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling que especifica o seguinte:

- A porcentagem do grupo a ser executado como instâncias sob demanda (`0`) e um número base de instâncias sob demanda com as quais começar (`1`)
- Os tipos de instância a serem iniciadas em ordem de prioridade (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)

- As sub-redes nas quais iniciar as instâncias (`subnet-5ea0c127`, `subnet-6194ea3b`, `subnet-c934b782`) Cada um corresponde a uma zona de disponibilidade diferente.
- O modelo de execução (`my-launch-template`) e a versão do modelo de execução (`$Default`)

Quando o Amazon EC2 Auto Scaling tenta atender à sua capacidade sob demanda, ele executa o tipo de instância `c5.large` primeiro. As instâncias spot vêm do grupo spot ideal em cada zona de disponibilidade com base na capacidade da instância spot.

JSON

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Default"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
```

```
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandBaseCapacity": 1,
            "OnDemandPercentageAboveBaseCapacity": 0,
            "SpotAllocationStrategy": "capacity-optimized"
        }
    },
    "MinSize": 1,
    "MaxSize": 5,
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Como alternativa, você pode usar o create-auto-scaling-groupcomando a seguir para criar o grupo Auto Scaling. Isso faz referência a um arquivo YAML como o único parâmetro para o grupo do Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

O arquivo `config.yaml` contém o conteúdo a seguir.

```
---
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
```

```
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandBaseCapacity: 1
    OnDemandPercentageAboveBaseCapacity: 0
    SpotAllocationStrategy: capacity-optimized
MinSize: 1
MaxSize: 5
DesiredCapacity: 3
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Exemplo 2: Iniciar instâncias spot usando a estratégia de alocação **`capacity-optimized-prioritized`**

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling que especifica o seguinte:

- A porcentagem do grupo a ser executado como instâncias sob demanda (`0`) e um número base de instâncias sob demanda com as quais começar (1)
- Os tipos de instância a serem iniciadas em ordem de prioridade (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)
- As sub-redes nas quais iniciar as instâncias (`subnet-5ea0c127`, `subnet-6194ea3b`, `subnet-c934b782`) Cada um corresponde a uma zona de disponibilidade diferente.
- O modelo de execução (`my-launch-template`) e a versão do modelo de execução (`$Latest`)

Quando o Amazon EC2 Auto Scaling tenta atender à sua capacidade sob demanda, ele executa o tipo de instância `c5.large` primeiro. Quando o Amazon EC2 Auto Scaling tenta atender sua capacidade spot, ele honra as prioridades de tipo de instância com base no melhor esforço. No entanto, ele otimiza primeiro a capacidade.

JSON

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-asg",
```

```
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandBaseCapacity": 1,
            "OnDemandPercentageAboveBaseCapacity": 0,
            "SpotAllocationStrategy": "capacity-optimized-prioritized"
        }
    },
    "MinSize": 1,
    "MaxSize": 5,
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

```
}
```

YAML

Como alternativa, você pode usar o create-auto-scaling-groupcomando a seguir para criar o grupo Auto Scaling. Isso faz referência a um arquivo YAML como o único parâmetro para o grupo do Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

O arquivo `config.yaml` contém o conteúdo a seguir.

```
---
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandBaseCapacity: 1
    OnDemandPercentageAboveBaseCapacity: 0
    SpotAllocationStrategy: capacity-optimized-prioritized
MinSize: 1
MaxSize: 5
DesiredCapacity: 3
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Exemplo 3: Iniciar instâncias spot usando a estratégia de alocação **`lowest-price`** diversificada em dois grupos

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling que especifica o seguinte:

- O percentual do grupo a ser iniciado como instâncias sob demanda (`50`) (Isso não especifica um número base de instâncias sob demanda para começar.)
- Os tipos de instância a serem iniciadas em ordem de prioridade (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)
- As sub-redes nas quais iniciar as instâncias (`subnet-5ea0c127`, `subnet-6194ea3b`, `subnet-c934b782`) Cada um corresponde a uma zona de disponibilidade diferente.
- O modelo de execução (`my-launch-template`) e a versão do modelo de execução (`$Latest`)

Quando o Amazon EC2 Auto Scaling tenta atender à sua capacidade sob demanda, ele executa o tipo de instância `c5.large` primeiro. Para sua capacidade spot, o Amazon EC2 Auto Scaling tenta iniciar as instâncias spot uniformemente nos dois grupos de menor preço em cada zona de disponibilidade.

JSON

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
```

```
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 50,
            "SpotAllocationStrategy": "lowest-price",
            "SpotInstancePools": 2
        }
    },
    "MinSize": 1,
    "MaxSize": 5,
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Como alternativa, você pode usar o create-auto-scaling-groupcomando a seguir para criar o grupo Auto Scaling. Isso faz referência a um arquivo YAML como o único parâmetro para o grupo do Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

O arquivo `config.yaml` contém o conteúdo a seguir.

```
---
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
```

```
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandPercentageAboveBaseCapacity: 50
    SpotAllocationStrategy: lowest-price
    SpotInstancePools: 2
MinSize: 1
MaxSize: 5
DesiredCapacity: 3
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Exemplo 4: Iniciar Instâncias spot usando a estratégia de alocação **`price-capacity-optimized`**

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling que especifica o seguinte:

- O percentual do grupo a ser iniciado como instâncias sob demanda (`30`) (Isso não especifica um número base de instâncias sob demanda para começar.)
- Os tipos de instância a serem iniciadas em ordem de prioridade (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)
- As sub-redes nas quais iniciar as instâncias (`subnet-5ea0c127`, `subnet-6194ea3b`, `subnet-c934b782`) Cada um corresponde a uma zona de disponibilidade diferente.
- O modelo de execução (`my-launch-template`) e a versão do modelo de execução (`$Latest`)

Quando o Amazon EC2 Auto Scaling tenta atender à sua capacidade sob demanda, ele executa o tipo de instância `c5.large` primeiro. Para sua capacidade spot, o Amazon EC2 Auto Scaling tenta iniciar as instâncias spot de pools de instâncias spot com o preço mais baixo possível, mas também com capacidade ideal para o número de instâncias que estão sendo executadas.

JSON

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 30,
            "SpotAllocationStrategy": "price-capacity-optimized"
        }
    },
    "MinSize": 1,
    "MaxSize": 5,
```

```
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Como alternativa, você pode usar o create-auto-scaling-groupcomando a seguir para criar o grupo Auto Scaling. Isso faz referência a um arquivo YAML como o único parâmetro para o grupo do Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

O arquivo `config.yaml` contém o conteúdo a seguir.

```
---
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandPercentageAboveBaseCapacity: 30
    SpotAllocationStrategy: price-capacity-optimized
MinSize: 1
MaxSize: 5
DesiredCapacity: 3
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Configurar um grupo do Auto Scaling para usar ponderação da instância

Ao usar vários tipos de instância, você pode especificar quantas unidades associar a cada tipo de instância e, em seguida, especificar a capacidade do seu grupo com a mesma unidade de medida. Essa opção de especificação de capacidade é conhecida como ponderações.

Por exemplo, digamos que você execute um aplicativo de computação intensiva que tenha melhor desempenho com pelo menos 8 V CPUs e 15 GiB de RAM. Se você usar `c5.2xlarge` como sua unidade base, qualquer um dos tipos de instância do EC2 a seguir atenderá às necessidades da aplicação.

Exemplo de tipos de instância

Tipo de instância	vCPU	Memória (GiB)
`c5.2xlarge`	8	16
`c5.4xlarge`	16	32
`c5.12xlarge`	48	96
`c5.18xlarge`	72	144
`c5.24xlarge`	96	192

Por padrão, todos os tipos de instância têm o mesmo peso, independentemente do tamanho. Em outras palavras, se o Amazon EC2 Auto Scaling iniciar um tipo de instância grande ou pequeno, cada instância será considerada na capacidade desejada do grupo do Auto Scaling.

No entanto, com a ponderação de instâncias, você atribui um valor numérico que especifica quantas unidades de capacidade devem ser associadas a cada tipo de instância. Por exemplo, se as instâncias tiverem tamanhos diferentes, uma instância `c5.2xlarge` poderá ter o peso 2, uma `c5.4xlarge` (que é duas vezes maior) poderá ter o peso 4 e assim por diante. Quando o Amazon EC2 Auto Scaling reduz a escala do grupo, esses pesos se traduzem no número de unidades que cada instância conta para a capacidade desejada.

Os pesos não alteram quais tipos de instância o Amazon EC2 Auto Scaling escolhe iniciar; em vez disso, as estratégias de alocação fazem isso. Para obter mais informações, consulte Estratégias de alocação para vários tipos de instância.

> **Important**
>
> Para configurar um grupo de Auto Scaling para atingir a capacidade desejada usando o número de v CPUs ou a quantidade de memória de cada tipo de instância, recomendamos usar a seleção de tipo de instância baseada em atributos. A configuração do parâmetro `DesiredCapacityType` especifica automaticamente o número de unidades a serem associadas a cada tipo de instância com base no valor que você definiu para esse parâmetro. Para obter mais informações, consulte Crie um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos.

Conteúdo

- Considerações
- Comportamentos de peso da instância
- Configurar um grupo do Auto Scaling para usar pesos
- Exemplo de preço spot por unidade hora

Considerações

Esta seção discute as principais considerações na implementação eficaz da ponderação de instâncias.

- Escolha alguns tipos de instância que atendam às necessidades de desempenho da sua aplicação. Decida a ponderação que cada tipo de instância deve ter em relação à capacidade desejada do seu grupo do Auto Scaling com base em seus recursos. Os pesos se aplicam a instâncias atuais e futuras no grupo.
- Evite grandes intervalos entre ponderações. Por exemplo, não recomendamos especificar um peso 1 para um tipo de instância quando o próximo tipo de instância maior tiver um peso 200. A diferença entre os pesos menores e maiores também não deve ser extrema. Diferenças extremas de ponderação podem afetar negativamente a otimização do custo-desempenho..
- Especifique a capacidade desejada do grupo em unidades, não em instâncias. Por exemplo, se os seus pesos forem baseados em vCPUs, será necessário especificar o número desejado, mínimo e máximo de núcleos que você quer.
- Defina seus pesos e a capacidade desejada de forma que a capacidade desejada seja pelo menos duas a três vezes maior do que o seu maior peso.

Observe o seguinte ao atualizar grupos existentes:

- Quando você adicionar ponderações a um grupo existente, inclua ponderações para todos os tipos de instância atualmente em uso.
- Ao atualizar pesos existentes, o Amazon EC2 Auto Scaling iniciará ou encerrará instâncias para alcançar sua capacidade desejada com base nos novos pesos.
- Se você remover um tipo de instância, as instâncias em execução desse tipo manterão sua última ponderação, mesmo que não estejam mais definidas.

Comportamentos de peso da instância

Quando você usa pesos de instância, o Amazon EC2 Auto Scaling se comporta da seguinte maneira:

- A capacidade atual será a capacidade desejada ou acima dela. A capacidade atual pode exceder a capacidade desejada se forem executadas instâncias que excedam as unidades de capacidade desejada restantes. Por exemplo, vamos supor que você especifique dois tipos de instância, `c5.2xlarge` e `c5.12xlarge`, e atribua pesos de instância de 2 para `c5.2xlarge` e 12 para `c5.12xlarge`. Se houver cinco unidades restantes para atender a capacidade desejada, e o Amazon EC2 Auto Scaling provisionar uma,`c5.12xlarge` a capacidade desejada será excedida em sete unidades.
- Ao iniciar instâncias, o Amazon EC2 Auto Scaling prioriza a distribuição da capacidade entre as zonas de disponibilidade e o respeito às estratégias de alocação em vez de exceder a capacidade desejada.
- O Amazon EC2 Auto Scaling pode ultrapassar o limite máximo de capacidade para manter o equilíbrio entre as zonas de disponibilidade usando suas estratégias de alocação preferenciais. O limite rígido imposto pelo Amazon EC2 Auto Scaling é um valor que é igual à sua capacidade desejada mais o seu maior peso.

Configurar um grupo do Auto Scaling para usar pesos

Você pode configurar um grupo do Auto Scaling para usar pesos, conforme mostrado nos exemplos AWS CLI a seguir. Para obter instruções sobre como usar o console, consulte Criar um grupo misto de instâncias escolhendo manualmente os tipos de instância.

Para configurar um novo grupo do Auto Scaling para usar pesos (AWS CLI)

Use o comando create-auto-scaling-group. Por exemplo, o comando a seguir cria um novo grupo do Auto Scaling e adiciona pesos especificando o seguinte:

- O percentual do grupo a ser iniciado como instâncias sob demanda (`0`)
- A estratégia de alocação para instâncias spot em cada zona de disponibilidade (`capacity-optimized`)
- Os tipos de instância a serem executados em ordem de prioridade (`m4.16xlarge`, `m5.24xlarge`)
- Os pesos das instâncias que correspondem à diferença de tamanho relativo (vCPUs) entre os tipos de instância (16,24)
- As sub-redes nas quais iniciar as instâncias (`subnet-5ea0c127`, `subnet-6194ea3b`, `subnet-c934b782`), cada uma correspondente a uma zona de disponibilidade diferente
- O modelo de execução (`my-launch-template`) e a versão do modelo de execução (`$Latest`)

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "m4.16xlarge",
                    "WeightedCapacity": "16"
                },
                {
                    "InstanceType": "m5.24xlarge",
                    "WeightedCapacity": "24"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 0,
            "SpotAllocationStrategy": "capacity-optimized"
        }
    },
```

```
    "MinSize": 160,
    "MaxSize": 720,
    "DesiredCapacity": 480,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
    "Tags": []
}
```

Para configurar um grupo do Auto Scaling existente para usar pesos (AWS CLI)

Use o comando update-auto-scaling-group. Por exemplo, o comando a seguir adiciona pesos a tipos de instância em um grupo do Auto Scaling existente especificando o seguinte:

- Os tipos de instância a serem executados em ordem de prioridade (`c5.18xlarge`, `c5.24xlarge`, `c5.2xlarge`, `c5.4xlarge`)
- Os pesos das instâncias que correspondem à diferença de tamanho relativo (vCPUs) entre os tipos de instância (18,, 242,4)
- A nova capacidade desejada aumentada, que é maior do que o maior peso

```
aws autoscaling update-auto-scaling-group --cli-input-json file://~/config.json
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-existing-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "Overrides": [
                {
                    "InstanceType": "c5.18xlarge",
                    "WeightedCapacity": "18"
                },
                {
                    "InstanceType": "c5.24xlarge",
                    "WeightedCapacity": "24"
                },
                {
                    "InstanceType": "c5.2xlarge",
                    "WeightedCapacity": "2"
                },
                {
```

```
                    "InstanceType": "c5.4xlarge",
                    "WeightedCapacity": "4"
                }
            ]
        }
    },
    "MinSize": 0,
    "MaxSize": 100,
    "DesiredCapacity": 100
}
```

Para verificar os pesos usando a linha de comando

Use um dos seguintes comandos:

- describe-auto-scaling-groups (AWS CLI)
- Obtenha- ASAuto ScalingGroup (AWS Tools for Windows PowerShell)

Exemplo de preço spot por unidade hora

A tabela a seguir compara o preço por hora das instâncias spot em diferentes zonas de disponibilidade no Leste dos EUA (Norte da Virgínia) com o preço das instâncias sob demanda na mesma região. Os preços mostrados são preços de exemplo e não os preços atuais. Estes são seus custos por hora de instância.

Exemplo: preços spot por hora de instância

Tipo de instância	us-east-1a	us-east-1b	us-east-1c	Definição de preço sob demanda
c5.2xlarge	0,180 USD	0,191 USD	0,170 USD	0,34 USD
c5.4xlarge	0,341 USD	0,361 USD	0,318 USD	0,68 USD
c5.12xlarge	0,779 USD	0,777 USD	0,777 USD	2,04 USD
c5.18xlarge	1,207 USD	1,475 USD	1,357 USD	3,06 USD

Tipo de instância	us-east-1a	us-east-1b	us-east-1c	Definição de preço sob demanda
c5.24xlarge	1,555 USD	1,555 USD	1,555 USD	4,08 USD

Com os pesos de instâncias, você pode avaliar seus custos com base no que você usa por unidade de hora. Você pode determinar o preço por hora dividindo seu preço para um tipo de instância pelo número de unidades que ele representa. Para instâncias sob demanda, o preço por hora ao implantar um tipo de instância é igual ao que é ao implantar um tamanho diferente do mesmo tipo de instância. Por outro lado, o preço spot por hora varia por grupo spot.

O exemplo a seguir mostra como o cálculo do preço spot por unidade de hora funciona com pesos de instância. Para facilitar o cálculo, digamos que você queira iniciar instâncias spot somente em `us-east-1a`. O preço unitário por hora está capturado na tabela a seguir.

Exemplo: preço spot por unidade hora

Tipo de instância	us-east-1a	Peso da instância	Preço por hora
c5.2xlarge	0,180 USD	2	0,090 USD
c5.4xlarge	0,341 USD	4	0,085 USD
c5.12xlarge	0,779 USD	12	0,065 USD
c5.18xlarge	1,207 USD	18	0,067 USD
c5.24xlarge	1,555 USD	24	0,065 USD

Usar vários modelos de execução

Além de usar vários tipos de instância, você também pode usar vários modelos de execução.

Por exemplo, digamos que você configure um grupo do Auto Scaling para aplicações de computação intensiva e queira incluir uma combinação de tipos de instância C5, C5a e C6g. No entanto, as instâncias C6g apresentam um processador AWS Graviton baseado na arquitetura Arm de 64

bits, enquanto as instâncias C5 e C5a são executadas em processadores Intel x86 de 64 bits. As instâncias AMIs for C5 e C5a funcionam em cada uma dessas instâncias, mas não nas instâncias C6g. Para resolver o problema, use um modelo de execução diferente para as instâncias C6g. Você ainda pode usar o mesmo modelo de execução para instâncias C5 e C5a.

Esta seção contém procedimentos para usar o AWS CLI para realizar tarefas relacionadas ao uso de vários modelos de execução. No momento, esse recurso estará disponível somente se você usar a AWS CLI ou um SDK, e não está disponível no console.

Conteúdo

- Configurar um grupo do Auto Scaling para usar vários modelos de execução
- Recursos relacionados

Configurar um grupo do Auto Scaling para usar vários modelos de execução

Você pode configurar um grupo do Auto Scaling para usar vários modelos de execução, conforme mostrado nos exemplos a seguir.

Para configurar um novo grupo do Auto Scaling para usar vários modelos de execução (AWS CLI)

Use o comando create-auto-scaling-group. Por exemplo, o comando a seguir cria um novo grupo do Auto Scaling. Ele especifica os tipos de instância `c5.large`, `c5a.large`, e `c6g.large` e define um novo modelo de execução para o tipo de instância `c6g.large` para garantir que uma AMI apropriada seja usada para iniciar instâncias Arm. O Amazon EC2 Auto Scaling usa a ordem de tipos de instâncias para determinar qual tipo de instância usar primeiro ao atender à capacidade sob demanda.

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
  "AutoScalingGroupName":"my-asg",
  "MixedInstancesPolicy":{
    "LaunchTemplate":{
      "LaunchTemplateSpecification":{
        "LaunchTemplateName":"my-launch-template-for-x86",
        "Version":"$Latest"
```

```
      },
      "Overrides":[
        {
          "InstanceType":"c6g.large",
          "LaunchTemplateSpecification": {
            "LaunchTemplateName": "my-launch-template-for-arm",
            "Version": "$Latest"
          }
        },
        {
          "InstanceType":"c5.large"
        },
        {
          "InstanceType":"c5a.large"
        }
      ]
    },
    "InstancesDistribution":{
      "OnDemandBaseCapacity": 1,
      "OnDemandPercentageAboveBaseCapacity": 50,
      "SpotAllocationStrategy": "capacity-optimized"
    }
  },
  "MinSize":1,
  "MaxSize":5,
  "DesiredCapacity":3,
  "VPCZoneIdentifier":"subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
  "Tags":[ ]
}
```

Para configurar um grupo do Auto Scaling existente para usar vários modelos de execução (AWS CLI)

Use o comando update-auto-scaling-group. Por exemplo, o comando a seguir atribui o modelo de execução chamado *my-launch-template-for-arm* ao tipo de instância *c6g.large* do grupo do Auto Scaling chamado *my-asg*.

```
aws autoscaling update-auto-scaling-group --cli-input-json file://~/config.json
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
```

```
  "AutoScalingGroupName":"my-asg",
  "MixedInstancesPolicy":{
    "LaunchTemplate":{
      "Overrides":[
        {
          "InstanceType":"c6g.large",
          "LaunchTemplateSpecification": {
            "LaunchTemplateName": "my-launch-template-for-arm",
            "Version": "$Latest"
          }
        },
        {
          "InstanceType":"c5.large"
        },
        {
          "InstanceType":"c5a.large"
        }
      ]
    }
  }
}
```

Para configurar um novo grupo do Auto Scaling para usar vários modelos de execução com seleção de tipo de instância baseada em atributos (AWS CLI)

Use o comando create-auto-scaling-group. Por exemplo, o comando a seguir cria um novo grupo de Auto Scaling especificando um modelo de execução para instâncias AWS Graviton com uma AMI ARM e um modelo de execução adicional para instâncias baseadas em AMD ou Intel com uma AMI x86. Em seguida, usa duas vezes a seleção de instância baseada em atributos para selecionar entre uma ampla variedade de tipos de instância para cada arquitetura de CPU. Você pode adicionar uma configuração semelhante a um grupo de Auto Scaling existente com o update-autoscaling-groupcomando.

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
  "AutoScalingGroupName":"my-asg",
  "MixedInstancesPolicy":{
    "LaunchTemplate":{
```

```
      "LaunchTemplateSpecification":{
        "LaunchTemplateName":"my-launch-template-for-arm",
        "Version":"$Latest"
      },
      "Overrides":[
        {
          "InstanceRequirements": {
            "VCpuCount": {"Min": 2},
            "MemoryMiB": {"Min": 2048},
            "CpuManufacturers": ["amazon-web-services"]
          }
         },
         {
           "InstanceRequirements": {
            "VCpuCount": {"Min": 2},
            "MemoryMiB": {"Min": 2048},
            "CpuManufacturers": ["intel", "amd"]
          },
          "LaunchTemplateSpecification": {
            "LaunchTemplateName": "my-launch-template-for-x86",
            "Version": "$Latest"
          }
         }
      ]
    },
    "InstancesDistribution":{
      "OnDemandPercentageAboveBaseCapacity": 0,
      "SpotAllocationStrategy": "price-capacity-optimized"
    }
  },
  "MinSize":1,
  "MaxSize":10,
  "DesiredCapacity":6,
  "VPCZoneIdentifier":"subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
  "Tags":[ ]
}
```

Para verificar os modelos de execução de um grupo do Auto Scaling

Use um dos seguintes comandos:

- describe-auto-scaling-groups (AWS CLI)
- Obtenha- ASAuto ScalingGroup (AWS Tools for Windows PowerShell)

Recursos relacionados

Você pode encontrar um exemplo de especificação de vários modelos de execução usando a seleção de tipo de instância baseada em atributos em um CloudFormation modelo em re:POST.AWS

Criar grupos do Auto Scaling usando configurações de execução

Important

Limitações:

- Desde 1º de janeiro de 2023 não há mais suporte aos novos tipos de instância do Amazon EC2 nas configurações de execução. Isso inclui suporte para qualquer tipo de instância adicionado Região da AWS após o lançamento inicial da região.
- As contas criadas em ou após 1º de junho de 2023 não podem criar novas configurações de execução usando o console.
- As contas criadas em ou após 1º de outubro de 2024 não podem criar novas configurações de lançamento usando nenhum método (console AWS CLI, API ou CloudFormation).

Migre para modelos de execução para garantir que você não precise criar novas configurações de execução agora ou no futuro. Para obter informações sobre a migração de grupos do Auto Scaling para modelos de execução, consulte Migre seus grupos do Auto Scaling para modelos de execução.

Se você criou uma configuração de execução ou uma instância do EC2, poderá criar um grupo do Auto Scaling que use uma configuração de execução como modelo de configuração para suas instâncias do EC2. A configuração de execução especifica informações como ID da AMI, tipo de instância, par de chaves, grupos de segurança e mapeamento de dispositivos de bloco para suas instâncias. Para obter informações sobre como criar configurações de inicialização, consulte Criar uma configuração de execução.

É necessário ter permissões suficientes para criar um grupo do Auto Scaling. Você também deve ter permissões suficientes para criar o perfil vinculado ao serviço que o Amazon EC2 Auto Scaling usa para realizar ações por sua própria conta se ela ainda não existir. Para ver exemplos de políticas do IAM que um administrador pode usar como referência para lhe conceder permissões, consulte Exemplos de políticas baseadas em identidade.

Conteúdo

- Criar um grupo do Auto Scaling usando uma configuração de execução
- Crie um grupo de Auto Scaling a partir de uma instância existente usando o AWS CLI

Criar um grupo do Auto Scaling usando uma configuração de execução

> **Important**
>
> Fornecemos informações sobre configurações de execução para clientes que ainda não migraram das configurações de execução para os modelos de execução. Para obter informações sobre como migrar seu grupo do Auto Scaling, para lançar modelos, consulte Migre seus grupos do Auto Scaling para modelos de execução.

Ao criar um grupo do Auto Scaling, será necessário especificar as informações necessárias para configurar as instâncias do Amazon EC2, as zonas de disponibilidade e sub-redes VPC para as instâncias, a capacidade desejada e os limites de capacidade mínimo e máximo.

O procedimento a seguir demonstra como criar um grupo do Auto Scaling usando uma configuração de execução. Não é possível modificar uma configuração de execução depois que ela é criada, mas você pode substituí-la por um grupo do Auto Scaling. Para obter mais informações, consulte Alterar a configuração de execução de um grupo do Auto Scaling.

Pré-requisitos

- É necessário ter criado uma configuração de execução. Para obter mais informações, consulte Criar uma configuração de execução.

Para criar um grupo do Auto Scaling usando uma configuração de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a mesma Região da AWS que você usou ao criar a configuração de inicialização.
3. Selecione Criar um grupo do Auto Scaling.

4. Na página Choose launch template or configuration (Escolher o modelo ou a configuração de execução), em Auto Scaling group name (Nome do grupo do Auto Scaling) insira um nome para o grupo do Auto Scaling.
5. Para escolher uma configuração de execução, faça o seguinte:

   a. Em Launch Template (Modelo de execução), selecione Switch to launch configuration (Alternar para configuração de execução).

   b. Em Launch configuration (Configuração de execução), escolha uma configuração de execução existente.

   c. Verifique se a configuração de execução oferece suporte a todas as opções que você está planejando usar e escolha Next (Próximo).
6. Na página Configurar opções de execução da instância, sob Rede, em VPC, escolha uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado na configuração de execução.
7. Em Zonas de disponibilidade e sub-redes, escolha uma ou mais sub-redes na VPC especificada. Use sub-redes em várias zonas de disponibilidade para alta disponibilidade. Para obter mais informações, consulte Considerações sobre a escolha de sub-redes da VPC.
8. Escolha Próximo.

   Ou é possível aceitar o restante dos padrões e escolher Skip to review (Avançar para análise).
9. (Opcional) Na página Configure advanced options (Configurar opções avançadas), configure as seguintes opções e escolha Next (Próximo):

   a. (Opcional) Em Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Amazon EBS. Para obter mais informações, consulte Monitore as instâncias do Auto Scaling com volumes do Amazon EBS com problemas usando verificações de integridade.

   b. Opcional em Tempo de carência da verificação de integridade, insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

   c. Em Configurações adicionais, Monitoramento, escolha se deseja ativar a coleta de métricas de CloudWatch grupo. Essas métricas fornecem medições que podem ser indicadores de um problema potencial, como número de instâncias de terminação ou número de instâncias

pendentes. Para obter mais informações, consulte Monitore CloudWatch métricas para seus grupos e instâncias do Auto Scaling.

d. Em Enable default instance warmup (Habilitar o aquecimento de instância padrão), selecione essa opção e escolha o tempo de aquecimento para sua aplicação. Se você estiver criando um grupo de Auto Scaling que tenha uma política de escalabilidade, o recurso padrão de aquecimento de instâncias melhora as CloudWatch métricas da Amazon usadas para escalabilidade dinâmica. Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

10. (Opcional) Na página Configure group size and scaling policies (Configurar o tamanho do grupo e as políticas de escalabilidade), configure as seguintes opções e escolha Next (Próximo):

    a. Sob o tamanho do grupo, para a capacidade desejada, insira o número inicial de instâncias para ser lançado.

    b. Na seção Escalabilidade, em Limites de escalabilidade, se o novo valor para Capacidade desejada for maior que Capacidade mínima desejada e Capacidade máxima desejada, a Capacidade máxima desejada será automaticamente aumentada para o novo valor da capacidade desejada. É possível alterar esses limites conforme necessário. Para obter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.

    c. Em Escalabilidade automática, escolha se você deseja criar uma política de escalabilidade de rastreamento de destino. Você também pode criar essa política depois de criar seu grupo do Auto Scaling.

       Se você escolher a política de escalabilidade de rastreamento de destino, siga as instruções em Criar uma política de dimensionamento com monitoramento do objetivo para criar a política.

    d. Em Política de manutenção de instâncias, escolha se você deseja criar uma política de manutenção de instâncias. Você também pode criar essa política depois de criar seu grupo do Auto Scaling. Siga as instruções Definir uma política de manutenção de instâncias para criar a política.

    e. Em Instance scale-in protection (Proteção de redução de instâncias), escolha se deseja habilitar a proteção de redução de instâncias. Para obter mais informações, consulte Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância.

11. (Opcional) Para receber notificações, em Add notification (Adicionar notificação), configure a notificação e, depois, escolha Next (Próximo). Para obter mais informações, consulte Configurar notificações do Amazon SNS para o Amazon EC2 Auto Scaling.

12. (Opcional) Para adicionar tags, escolha Add tag (Adicionar tag), forneça uma chave e um valor para cada tag e, depois, escolha Next (Próximo). Para obter mais informações, consulte Etiquetar grupos e instâncias do Auto Scaling.
13. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Para criar um grupo do Auto Scaling usando a linha de comando

Você pode usar um dos comandos a seguir:

- create-auto-scaling-group (AWS CLI)
- Novo- ASAuto ScalingGroup (AWS Tools for Windows PowerShell)

Crie um grupo de Auto Scaling a partir de uma instância existente usando o AWS CLI

> **Important**
>
> Fornecemos informações sobre configurações de execução para clientes que ainda não migraram das configurações de execução para os modelos de execução. Para obter informações sobre a migração de grupos do Auto Scaling para modelos de execução, consulte Migre seus grupos do Auto Scaling para modelos de execução.

Se esta for a primeira vez que você cria um grupo do Auto Scaling, recomendamos que você use o console para criar um modelo de execução a partir de uma instância do EC2 existente. Em seguida, use o modelo de execução para criar um novo grupo do Auto Scaling. Para esse procedimento, consulte Criar um grupo do Auto Scaling usando o assistente de execução do Amazon EC2.

O procedimento a seguir mostra como criar um grupo do Auto Scaling especificando uma instância existente a ser usada como base para iniciar outras instâncias. Vários parâmetros são necessárias para criar uma instância do EC2, como o ID do imagem de máquina da Amazon (AMI), o tipo de instância, o par de chaves e o grupo de segurança. Todas essas informações também são usadas pelo Amazon EC2 Auto Scaling para iniciar instâncias em seu nome quando houver necessidade de escalar. Essas informações são armazenadas em um modelo de execução ou uma configuração de execução.

Quando você usa uma instância existente, o Amazon EC2 Auto Scaling cria um grupo do Auto Scaling que inicia instâncias com base em uma configuração de execução criada ao mesmo tempo. A nova configuração de execução tem o mesmo nome do grupo do Auto Scaling e inclui determinados detalhes de configuração da instância identificada.

Os detalhes de configuração a seguir são copiados da instância identificada para a configuração de execução:

- ID DA AMI
- Tipo de instância
- Par de chaves
- Grupos de segurança
- Tipo de endereço IP (público ou privado)
- Perfil da instância do IAM, se aplicável
- Monitoramento (verdadeiro ou falso)
- Otimizado para o EBS (verdadeiro ou falso)
- Configuração de locação, se executando dentro de uma VPC (compartilhada ou dedicada)
- ID do kernel e ID do disco RAM, se aplicável
- Dados do usuário, se especificado
- Preço (máximo) do spot

A sub-rede VPC e a zona de disponibilidade são copiadas da instância identificada para a própria definição de recursos do grupo do Auto Scaling.

Se a instância identificada estiver em um grupo de posicionamento, o novo grupo do Auto Scaling iniciará instâncias no mesmo grupo de posicionamento da instância identificada. Como as configurações de execução não permitem que um grupo de posicionamento seja especificado, o grupo de posicionamento é copiado para o atributo `PlacementGroup` do novo grupo do Auto Scaling.

Os seguintes detalhes da configuração não são copiados da instância identificada:

- Armazenamento: os dispositivos de bloco (volumes do EBS e volumes de armazenamento de instâncias) não são copiados da instância identificada. Em vez disso, o mapeamento de dispositivos de bloco criado como parte da criação da AMI determina quais dispositivos são usados.

- Número de interfaces de rede: as interfaces de rede não são copiadas da instância identificada. Em vez disso, o Amazon EC2 Auto Scaling usa suas configurações padrão para criar uma interface de rede, que é a interface de rede primária (eth0).
- Opções de metadados da instância: as configurações acessíveis de metadados, versão de metadados e limite de salto de resposta de token não são copiadas da instância identificada. Em vez disso, o Amazon EC2 Auto Scaling usa suas configurações padrão. Para obter mais informações, consulte Configurar as opções de metadados da instância.
- Balanceadores de carga: se a instância identificada estiver registrada em um ou mais balanceadores de carga, as informações sobre o balanceador de carga não serão copiadas para o balanceador de carga nem no atributo do grupo de destino do novo grupo do Auto Scaling.
- Etiquetas: se a instância identificada tiver etiquetas, elas não serão copiadas para o atributo de `Tags` do novo grupo do Auto Scaling.

Pré-requisitos

A instância do EC2 deve atender aos seguintes critérios:

- A instância não é um membro de outro grupo do Auto Scaling.
- A instância está no estado `running`.
- A AMI usada para iniciar a instância ainda deve existir.

Crie um grupo do Auto Scaling a partir de uma instância do EC2 (AWS CLI).

O procedimento a seguir mostra como usar um comando CLI para criar um grupo do Auto Scaling a partir de uma instância do EC2.

Esse procedimento não adiciona a instância ao grupo do Auto Scaling. Para que a instância seja anexada, é necessário executar o comando attach-instances após a criação do grupo do Auto Scaling.

Antes de começar, localize o ID da instância do EC2 usando o console do Amazon EC2 ou o comando describe-instances.

Para usar a instância atual como modelo

- Use o create-auto-scaling-groupcomando a seguir para criar um grupo de Auto Scaling,`my-asg-from-instance`, a partir da instância do EC2. `i-123456789abcdefg0`

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg-from-instance \
  --instance-id i-123456789abcdefg0 --min-size 1 --max-size 2 --desired-capacity 2
```

Para verificar se seu grupo do Auto Scaling executou instâncias

- Use o describe-auto-scaling-groupscomando a seguir para verificar se o grupo Auto Scaling foi criado com êxito.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg-from-instance
```

O exemplo de reposta a seguir mostra que a capacidade desejada do grupo é 2, o grupo tem 2 instâncias em execução e a configuração de execução é chamada `my-asg-from-instance`.


```
{
  "AutoScalingGroups":[
    {
      "AutoScalingGroupName":"my-asg-from-instance",
      "AutoScalingGroupARN":"arn",
      "LaunchConfigurationName":"my-asg-from-instance",
      "MinSize":1,
      "MaxSize":2,
      "DesiredCapacity":2,
      "DefaultCooldown":300,
      "AvailabilityZones":[
        "us-west-2a"
      ],
      "LoadBalancerNames":[],
      "TargetGroupARNs":[],
      "HealthCheckType":"EC2",
      "HealthCheckGracePeriod":0,
      "Instances":[
        {
          "InstanceId":"i-34567890abcdef012",
          "InstanceType":"t2.micro",
          "AvailabilityZone":"us-west-2a",
          "LifecycleState":"InService",
          "HealthStatus":"Healthy",
          "LaunchConfigurationName":"my-asg-from-instance",
```

```
          "ProtectedFromScaleIn":false
        },
        {
          "InstanceId":"i-012345abcdefg6789",
          "InstanceType":"t2.micro",
          "AvailabilityZone":"us-west-2a",
          "LifecycleState":"InService",
          "HealthStatus":"Healthy",
          "LaunchConfigurationName":"my-asg-from-instance",
          "ProtectedFromScaleIn":false
        }
      ],
      "CreatedTime":"2020-10-28T02:39:22.152Z",
      "SuspendedProcesses":[ ],
      "VPCZoneIdentifier":"subnet-0abc1234",
      "EnabledMetrics":[ ],
      "Tags":[ ],
      "TerminationPolicies":[
        "Default"
      ],
      "NewInstancesProtectedFromScaleIn":false,
      "ServiceLinkedRoleARN":"arn",
      "TrafficSources":[]
    }
  ]
}
```

Para visualizar a configuração de execução

- Use o describe-launch-configurationscomando a seguir para ver os detalhes da configuração de inicialização.

```
aws autoscaling describe-launch-configurations --launch-configuration-names my-asg-from-instance
```

A seguir está um exemplo de saída:

```
{
  "LaunchConfigurations":[
    {
      "LaunchConfigurationName":"my-asg-from-instance",
```

```
      "LaunchConfigurationARN":"arn",
      "ImageId":"ami-234567890abcdefgh",
      "KeyName":"my-key-pair-uswest2",
      "SecurityGroups":[
        "sg-12abcdefgh3456789"
      ],
      "ClassicLinkVPCSecurityGroups":[ ],
      "UserData":"",
      "InstanceType":"t2.micro",
      "KernelId":"",
      "RamdiskId":"",
      "BlockDeviceMappings":[ ],
      "InstanceMonitoring":{
        "Enabled":true
      },
      "CreatedTime":"2020-10-28T02:39:22.321Z",
      "EbsOptimized":false,
      "AssociatePublicIpAddress":true
    }
  ]
}
```

Para terminar as instâncias

- Você pode terminar a instância se não precisar mais dela. O seguinte comando terminate-instances termina a instância `i-123456789abcdefg0`.

  ```
  aws ec2 terminate-instances --instance-ids i-123456789abcdefg0
  ```

  Depois de terminar uma instância do Amazon EC2, você não poderá reiniciar a instância. Depois do término, seus dados são excluídos e o volume não pode mais ser conectado a nenhuma instância. Para saber mais sobre como terminar instâncias, consulte Terminate an instance (Como terminar uma instância) no Guia do usuário do Amazon EC2 para instâncias do Linux.

Execute instâncias de forma síncrona

O Amazon EC2 Auto Scaling fornece dois métodos para iniciar instâncias em seu grupo de Auto Scaling: comportamento de escalabilidade assíncrono e provisionamento síncrono usando a API. LaunchInstances

Com o provisionamento síncrono, você usa a LaunchInstances API para solicitar um número específico de instâncias em uma zona de disponibilidade específica. O provisionamento síncrono oferece os seguintes benefícios:

- Feedback imediato sobre a disponibilidade de capacidade em zonas de disponibilidade específicas
- Controle preciso sobre quais instâncias da Zona de Disponibilidade são executadas em
- Instância determinística IDs para uso imediato em sistemas de orquestração
- Decisões de escalabilidade em tempo real com base nas restrições reais de capacidade
- Dimensionamento mais rápido, eliminando os tempos de espera para lançamentos assíncronos do Auto Scaling

Com o Auto Scaling assíncrono, quando você altera a capacidade desejada ou quando uma política de escalabilidade é acionada, o Amazon EC2 Auto Scaling processa a solicitação de escalabilidade e inicia instâncias em segundo plano. Você deve monitorar as atividades de escalabilidade ou descrever seu grupo de Auto Scaling para determinar quando as instâncias são lançadas com sucesso.

> **Note**
>
> - A LaunchInstances API só funciona com grupos de Auto Scaling que usam modelos de lançamento. Grupos de Auto Scaling que usam configurações de inicialização não são suportados. Se seu grupo de Auto Scaling usa uma configuração de execução, você deve migrar para um modelo de execução antes de usar o provisionamento síncrono.
> - A LaunchInstances API oferece suporte a políticas de instâncias mistas somente com opções de compra totalmente sob demanda ou totalmente à vista. Políticas mistas combinando instâncias sob demanda e spot não são suportadas.
> - Para grupos de Auto Scaling que abrangem várias zonas de disponibilidade, você deve especificar a zona de disponibilidade ou sub-rede de destino. Para grupos Single-AZ, esse parâmetro é opcional.

Provisionamento síncrono e escalabilidade assíncrona

Provisionamento síncrono

Quando você usa a LaunchInstances API, o Amazon EC2 Auto Scaling:

- Tenta iniciar imediatamente as instâncias solicitadas usando CreateFleet
- Espera para retornar CreateFleet a instância IDs antes de responder
- Retorna informações sobre o sucesso da instância IDs, dos tipos de instância e da zona de disponibilidade
- Retorna códigos de erro específicos e detalhes sobre falhas
- Fornece feedback imediato, permitindo decisões de escalonamento em tempo real

Escalabilidade assíncrona

Quando você usa métodos assíncronos de Auto Scaling, como alterar a capacidade desejada ou usar políticas de escalabilidade, o Amazon EC2 Auto Scaling:

- Atualiza a capacidade desejada na API, mas não retornará as instâncias imediatamente
- A instância de planos é iniciada automaticamente em todas as zonas de disponibilidade
- Lança instâncias por meio de fluxos de trabalho em segundo plano
- Distribui automaticamente a capacidade em várias zonas de disponibilidade para equilibrar
- Lida com falhas de lançamento com lógica de repetição integrada

Você deve pesquisar as atividades de escalonamento ou descrever seu grupo de Auto Scaling para verificar o status das operações de lançamento.

Limitações e considerações

Ao trabalhar com o provisionamento síncrono, lembre-se das seguintes notas e limitações:

- Estado da instância após o lançamento — As instâncias retornadas pela API estão em estado pendente. Eles ainda podem falhar durante processos de fluxo de trabalho subsequentes ou ganchos do ciclo de vida. Uma resposta de API bem-sucedida significa que o EC2 aceitou a solicitação de execução e retornou os IDs da instância. As instâncias não são automaticamente consideradas totalmente prontas para cargas de trabalho e devem concluir os processos padrão do ciclo de vida do EC2 e do Auto Scaling.
- Limitação de piscinas aquecidas — atualmente não há suporte para grupos de Auto Scaling com piscinas aquecidas. Se você tentar chamar a LaunchInstances API em um grupo de Auto Scaling que tenha um pool de aquecimento configurado, a API executará uma inicialização a frio em vez de usar instâncias de pool de aquecimento e retornará um UnsupportedOperation erro. Para obter mais informações sobre partidas a frio, consulte Limitações de piscinas aquecidas.

- Tempo limite da API e novas tentativas — se a CreateFleet operação subjacente demorar mais do que o esperado, a API poderá atingir o tempo limite e retornar um token de idempotência. Você pode tentar usar o mesmo novamente ClientToken para rastrear a operação de inicialização original ou usar describe-instances com o token do cliente para verificar as instâncias iniciadas.
- Restrições da zona de disponibilidade — Se seu grupo de Auto Scaling abrange várias zonas de disponibilidade e tem o rebalanceamento da zona de disponibilidade ativado, iniciar instâncias de forma síncrona pode causar conflitos operacionais:
  - Limitação única de AZ por chamada — Cada chamada de LaunchInstances API só pode atingir uma zona de disponibilidade, mesmo que seu grupo de Auto Scaling se estenda por várias zonas.
  - Conflitos de rebalanceamento de AZ - Se seu grupo de Auto Scaling tiver o rebalanceamento de AZ ativado, chamadas sequenciais entre AZs diferentes podem acionar lançamentos assíncronos adicionais, resultando em mais instâncias do que o pretendido. Considere suspender o rebalanceamento do AZ para um controle preciso da capacidade. Para obter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.
- Cenários de sucesso parcial — A `LaunchInstances` API pode retornar sucesso parcial se apenas parte da capacidade solicitada estiver disponível, o que é o comportamento normal do EC2. A API retorna instâncias iniciadas com sucesso junto com detalhes de erros de execuções malsucedidas. Para casos de uso que exigem que todas as instâncias sejam executadas juntas (como aplicativos que precisam de todas as instâncias na mesma AZ para baixa latência), você precisará encerrar instâncias parcialmente iniciadas e tentar novamente em uma AZ diferente. Considere esse comportamento ao criar uma lógica de repetição para cargas de trabalho sensíveis à capacidade.
- Pesos de instância — Se seu grupo de Auto Scaling usa pesos de instância, RequestedCapacity o parâmetro representa unidades de capacidade ponderada, não o número de instâncias. O número real de instâncias lançadas depende dos tipos de instância selecionados e de seus pesos configurados. O EC2 Auto Scaling limita o lançamento a 100 instâncias por chamada de API, independentemente da capacidade ponderada solicitada.
- Tipos de instâncias mistas — A LaunchInstances API usa a política de instâncias mistas existente do seu grupo de Auto Scaling para determinar quais tipos de instância iniciar. A API inicia instâncias de acordo com a estratégia de alocação e as prioridades do tipo de instância do seu grupo.

Lançamento de instâncias com provisionamento síncrono

Você pode usar a LaunchInstances API para executar de forma síncrona um número específico de instâncias em seu grupo de Auto Scaling. A API inicia instâncias na zona de disponibilidade ou sub-rede que você especifica e retorna imediatamente as informações da instância IDs ou do erro.

Pré-requisitos

Antes de usar a LaunchInstances API, você deve ter:

- Um grupo de Auto Scaling que usa um modelo de execução (configurações de execução não são suportadas)
- Você precisa ter permissões para as seguintes ações do IAM:
  - `autoscaling:LaunchInstances`
  - `ec2:CreateFleet`
  - `ec2:DescribeLaunchTemplateVersions`

Execute instâncias com provisionamento síncrono

Você pode iniciar instâncias com provisionamento síncrono por meio do. AWS CLI

AWS CLI

Para iniciar instâncias com provisionamento síncrono:

```
aws autoscaling launch-instances \
        --auto-scaling-group-name group-name \
        --requested-capacity number \
        [--availability-zones zone-name] \
        [--subnet-ids subnet-id] \
        [--availability-zone-ids zone-id] \
        [--retry-strategy none|retry-with-group-configuration] \
        [--client-token token]
```

Exemplos

Lançamento de instâncias em uma zona de disponibilidade específica

```
aws autoscaling launch-instances \
```

```
        --auto-scaling-group-name my-asg \
        --requested-capacity 3 \
        --availability-zones us-east-1a \
        --retry-strategy retry-with-group-configuration
```

Lançamento de instâncias em uma sub-rede específica

```
aws autoscaling launch-instances \
        --auto-scaling-group-name my-asg \
        --requested-capacity 2 \
        --subnet-ids subnet-12345678 \
        --retry-strategy none \
        --client-token my-unique-token-123
```

Lidando com respostas

Exemplo de uma resposta bem-sucedida:

```
{
    "AutoScalingGroupName": "my-asg",
    "ClientToken": "my-unique-token-123",
    "Instances": [
        {
            "InstanceType": "m5.xlarge",
            "AvailabilityZone": "us-east-1a",
            "AvailabilityZoneId": "use1-az1",
            "SubnetId": "subnet-12345678",
            "MarketType": "OnDemand",
            "InstanceIds": ["i-0123456789abcdef0", "i-0fedcba9876543210"]
        }
    ],
    "Errors": []
}
```

Exemplo de uma resposta com erros

```
{
    "AutoScalingGroupName": "my-asg",
    "ClientToken": "my-unique-token-123",
```

```
    "Instances": [],
    "Errors": [
       {
        "InstanceType": "m5.large",
        "AvailabilityZone": "us-east-1a",
        "AvailabilityZoneId": "use1-az1",
        "SubnetId": "subnet-12345678",
        "MarketType": "OnDemand",
        "ErrorCode": "InsufficientInstanceCapacity",
        "ErrorMessage": "There is not enough capacity to fulfill your request for
 instance type 'm5.large' in 'us-east-1a'"
        }
    ]
}
```

Lidar com falhas de lançamento e novas tentativas

Quando a LaunchInstances API encontra falhas, você pode implementar estratégias de repetição usando tokens de idempotência e políticas de repetição apropriadas.

Você pode usar o parâmetro client-token para repetir as solicitações. Você também pode usar as seguintes estratégias de repetição:

- `RetryStrategy: none`(padrão) - Se a chamada da API falhar, a capacidade desejada do grupo Auto Scaling permanecerá inalterada e nenhuma nova tentativa automática ocorrerá.
- `RetryStrategy: retry-with-group-configuration`- Se a chamada de API falhar, a capacidade desejada do grupo de Auto Scaling será aumentada na quantidade solicitada, e o Auto Scaling tentará automaticamente iniciar instâncias novamente usando a configuração e os processos padrão do grupo.

O comportamento de repetição `RetryStrategy: retry-with-group-configuration` depende do tipo de falha:

- Erros de validação: a capacidade desejada não é aumentada, pois a operação não pode continuar. Por exemplo, parâmetros inválidos ou configurações não suportadas.
- Erros de capacidade: a capacidade desejada é aumentada e o Auto Scaling tentará executar novamente as instâncias de forma assíncrona usando os processos normais de escalabilidade do grupo.

Usando tokens de cliente para idempotência

O `client-token` parâmetro garante operações idempotentes e permite novas tentativas seguras de solicitações de lançamento.

Comportamentos principais:

- Os tokens do cliente têm uma vida útil de 8 horas a partir da solicitação inicial
- Tentar novamente com o mesmo token de cliente em 8 horas retorna a resposta em cache em vez de iniciar novas instâncias
- Após 8 horas, o mesmo token de cliente iniciará uma nova operação de lançamento

Update an Auto Scaling group

Você pode atualizar a maioria dos detalhes do grupo do Auto Scaling. Você não pode atualizar o nome de um grupo do Auto Scaling nem alterá-lo. Região da AWS

Para atualizar um grupo do Auto Scaling (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Escolha seu grupo do Auto Scaling para exibir informações sobre o grupo, com guias para Detalhes, Atividade, Escalabilidade automática, Gerenciamento de instâncias, Monitoramento e Atualização de Instâncias.
3. Escolha as guias das áreas de configuração de seu interesse e atualize as configurações conforme necessário. Para cada configuração que você editar, escolha Atualizar para salvar suas alterações na configuração do grupo do Auto Scaling.

   - Guia Detalhes

     Estas são as configurações gerais do seu grupo do Auto Scaling. Você pode editá-las e gerenciá-las da mesma forma que o fez durante a criação do grupo do Auto Scaling.

     A seção Configurações avançadas tem algumas opções que não estão disponíveis ao criar o grupo, como políticas de encerramento, tempo de espera, processos suspensos e vida útil máxima da instância. Você também pode exibir, mas não editar, o grupo de posicionamento e o perfil vinculado ao serviço do grupo do Auto Scaling.
   - Guia Integrações

    - Balanceamento de carga: Elastic Load Balancing

      Se o grupo estiver associado aos recursos do Elastic Load Balancing, consulte Adicione uma Zona de disponibilidade antes de alterar as zonas de disponibilidade. Algumas restrições no balanceador de carga podem impedir que você aplique alterações nas zonas de disponibilidade do seu grupo às zonas de disponibilidade do balanceador de carga.
    - Opções de integração do VPC Lattice: VPC Lattice
    - Mudança de zona de ARC: mudança de zona do grupo do Auto Scaling
- Guia de escalonamento automático
    - Políticas de escalabilidade dinâmica - Políticas de escalabilidade dinâmica
    - Políticas de escalabilidade preditiva - Políticas de escalabilidade preditiva
    - Ações programadas - Ações agendadas
- Guia Gerenciamento de instâncias
    - Ganchos do ciclo de vida - Ganchos do ciclo de vida
    - Pool ativo - Pools ativos
- Guia Atividades
    - Notificações de atividades - Notificações do Amazon SNS
- Guia Monitoramento
    - Há apenas uma única opção nessa guia, que permite ativar ou desativar a coleta de métricas de CloudWatch grupo.

Para atualizar um grupo do Auto Scaling usando a linha de comando

Você pode usar um dos comandos a seguir:

- update-auto-scaling-group (AWS CLI)
- Atualizar- ASAuto ScalingGroup (AWS Tools for Windows PowerShell)

Atualizar instâncias do Auto Scaling

Se você associar um novo modelo de execução ou configuração de execução a um grupo do Auto Scaling, todas as novas instâncias terão a configuração atualizada. As instâncias existentes continuam a ser executadas com a configuração com a qual foram originalmente iniciadas. Para aplicar as alterações às instâncias existentes, você tem as seguintes opções:

- Inicie uma atualização de instância para substituir as instâncias mais antigas. Para obter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling.
- Aguarde que as atividades de escalabilidade substituam gradualmente as instâncias mais antigas por instâncias mais novas com base nas suas políticas de encerramento.
- Encerre-as manualmente para que sejam substituídas pelo seu grupo do Auto Scaling.

> **Note**
>
> Você pode alterar os seguintes atributos de instância especificando-os como parte do modelo de execução ou da configuração de execução:
>
> - Imagem de máquina da Amazon (AMI)
> - dispositivos de blocos
> - key pair (par de chaves)
> - instance type (tipo de instância)
> - grupos de segurança
> - dados do usuário
> - monitoramento
> - Perfil de instância do IAM
> - Locação de localização
> - kernel
> - disco ram
> - Indica se a instância tem um endereço IP público.
> - a estratégia de distribuição da zona de disponibilidade

Estratégia de alocação de grupos do Auto Scaling e alterações de capacidade

Quando você altera uma estratégia de alocação de grupos do Auto Scaling, as instâncias existentes não são substituídas. Todas as novas instâncias iniciadas devido a eventos de aumento na escala horizontal seguirão a nova estratégia de alocação. Qualquer evento futuro de redução de escala na horizontal seguirá a política de encerramento e usará a nova estratégia de alocação se a política de

encerramento estiver definida como `Default` ou `AllocationStrategy`. Por exemplo, se você alterar a estratégia de alocação de `lowest-price` para `price-capacity-optimized`, talvez não haja encerramentos de instâncias, mas todas as novas instâncias serão lançadas com a nova estratégia de alocação. As alterações no tipo de instância não afetam as instâncias existentes.

Quando você altera determinados parâmetros, como o OnDemandBaseCapacityou o OnDemandPercentageAboveBaseCapacity, o Auto Scaling se reequilibra automaticamente se a porcentagem de instâncias sob demanda e instâncias spot não corresponder às novas especificações. Por exemplo, suponha que um grupo do Auto Scaling tenha a `OnDemandPercentageAboveBaseCapacity` definida como 50% de instâncias sob demanda e 50% de instâncias spot. Em seguida, a `OnDemandPercentageAboveBaseCapacity` é aumentada para 100% de instâncias sob demanda. O grupo do Auto Scaling se reequilibrará proativamente lançando novas instâncias sob demanda e encerrando instâncias spot. A política de manutenção da instância que você definiu determinará a ordem das atividades de execução e encerramento.

Etiquetar grupos e instâncias do Auto Scaling

Uma tag é um rótulo de atributo personalizado que você atribui ou AWS atribui a um AWS recurso. Cada tag da tem duas partes:

- Uma chave de etiqueta (por exemplo, `costcenter`, `environment` ou `project`)
- Um campo opcional conhecido como um valor de etiqueta (por exemplo, `111122223333` ou `production`)

As tags ajudam a:

- Acompanhe seus AWS custos. Você ativa essas tags no Gerenciamento de Faturamento e Custos da AWS painel. AWS usa as tags para categorizar seus custos e entregar um relatório mensal de alocação de custos para você. Para obter mais informações, consulte Uso de tags de alocação de custos no Guia do usuário do AWS Billing .
- Controle o acesso a grupos do Auto Scaling com base em tags. É possível usar condições em suas políticas do IAM para controlar o acesso aos grupos do Auto Scaling com base nas tags desse grupo. Para obter mais informações, consulte Etiquetas para segurança.
- Filtre e pesquise por grupos do Auto Scaling com base nas tags adicionadas. Para obter mais informações, consulte Usar etiquetas para filtrar grupos do Auto Scaling.

- Identifique e organize seus AWS recursos. Muitos Serviços da AWS oferecem suporte à marcação, então você pode atribuir a mesma tag a recursos de serviços diferentes para indicar que os recursos estão relacionados.

Você pode marcar grupos do Auto Scaling novos ou existentes. Você também pode propagar tags de um grupo do Auto Scaling para as instâncias do EC2 que ele executa.

As tags não são propagadas para volumes do Amazon EBS. Para adicionar tags a volumes do Amazon EBS, especifique as tags em um modelo de execução. Para obter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.

Você pode criar e gerenciar tags por meio do Console de gerenciamento da AWS, AWS CLI, ou SDKs.

Conteúdo

- Restrições de nomeação e uso de tags
- Ciclo de vida de marcação de instâncias do EC2
- Marcar seus grupos do Auto Scaling
- Excluir tags
- Etiquetas para segurança
- Controlar o acesso usando etiquetas
- Usar etiquetas para filtrar grupos do Auto Scaling

Restrições de nomeação e uso de tags

As restrições básicas a seguir se aplicam a tags:

- O número máximo de tags por recurso é 50.
- O número máximo de tags que você pode adicionar ou remover usando uma única chamada é 25.
- O comprimento máximo da chave é 128 caracteres Unicode.
- O número máximo de tags que você pode atribuir a um recurso é 256.
- As chaves e os valores de marcas diferenciam maiúsculas de minúsculas. Como melhor prática, adote uma estratégia para letras maiúsculas em tags e implemente-a de forma consistente em todos os tipos de recursos.

- Não use o `aws:` prefixo nos nomes ou valores de suas tags, pois ele está reservado para AWS uso. Você não pode editar nem excluir nomes ou valores de tags com esse prefixo, e elas não são contadas em sua quota de tags por recurso.

Ciclo de vida de marcação de instâncias do EC2

Se você tiver optado por propagar tags para suas instâncias do EC2, as tags serão gerenciadas da seguinte forma:

- Quando um grupo do Auto Scaling executa instâncias, ele adiciona tags às instâncias durante a criação do recurso, e não após o recurso ser criado.
- O grupo do Auto Scaling adiciona automaticamente uma etiqueta às instâncias com uma chave do `aws:autoscaling:groupName` e um valor do nome do grupo do Auto Scaling.
- Se você especificar tags de instância em seu modelo de execução e optar por propagar tags de seu grupo para suas instâncias, todas as tags serão mescladas. Se a mesma chave da etiqueta for especificada para uma etiqueta no modelo de execução e uma etiqueta no grupo do Auto Scaling, então, o valor da etiqueta do grupo terá precedência.
- Quando você anexa instâncias existentes, o grupo do Auto Scaling adiciona as tags às instâncias substituindo todas as tags existentes pela mesma chave de tag. Ele também adiciona uma etiqueta com uma chave do `aws:autoscaling:groupName` e um valor do nome do grupo do Auto Scaling.
- Quando você desvincula uma instância de um grupo do Auto Scaling, ele remove apenas a tag `aws:autoscaling:groupName`.

Marcar seus grupos do Auto Scaling

Quando você adiciona uma tag a seu grupo do Auto Scaling, você pode especificar se ela deve ser adicionada às instâncias iniciadas no grupo do Auto Scaling. Se você modificar uma tag, a versão atualizada da tag será adicionada às instâncias executadas no grupo do Auto Scaling depois da alteração. Se você criar ou modificar uma tag em um grupo do Auto Scaling, essas alterações não serão feitas em instâncias que já estão em execução no grupo do Auto Scaling.

Conteúdo

- Adicionar ou modificar tags (console)
- Adicionar ou modificar tags (AWS CLI)

Adicionar ou modificar tags (console)

Para marcar um grupo do Auto Scaling na criação

Ao usar o console do Amazon EC2 para criar um grupo do Auto Scaling, você pode especificar valores e chaves de tags na página Add tags (Configurar tags) do assistente de criação de grupo do Auto Scaling. Para propagar uma tag às instâncias executadas no grupo do Auto Scaling, mantenha a opção Tag New Instances (Marcar novas instâncias) para essa tag selecionada. Caso contrário, desmarque-a.

Para adicionar ou modificar tags de um grupo do Auto Scaling existente

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Detalhes escolha Tags, Editar.
4. Para modificar as tags existentes, edite Chave e Valor.
5. Para adicionar uma nova tag, escolha Adicionar tag e edite Chave e Valor. É possível manter a opção Tag new instances (Marcar novas instâncias) selecionada para adicionar a tag às instâncias executadas no grupo do Auto Scaling automaticamente e, caso contrário, desmarcá-la.
6. Ao concluir a inclusão de tags, selecione Update (Atualizar).

Adicionar ou modificar tags (AWS CLI)

Os exemplos a seguir mostram como usar o para adicionar tags AWS CLI ao criar grupos de Auto Scaling e para adicionar ou modificar tags para grupos de Auto Scaling existentes.

Para marcar um grupo do Auto Scaling na criação

Use o create-auto-scaling-groupcomando para criar um novo grupo de Auto Scaling e adicionar uma tag, por exemplo**`environment=production`**, ao grupo Auto Scaling. A tag também é adicionada a todas as instâncias executadas no grupo do Auto Scaling.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
```

```
  --launch-configuration-name my-launch-config --min-size 1 --max-size 3 \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --tags Key=environment,Value=production,PropagateAtLaunch=true
```

Para criar ou modificar tags de um grupo do Auto Scaling existente

Use o comando create-or-update-tags para criar ou modificar uma tag. Por exemplo, o comando a seguir adiciona as tags **costcenter=cc123** e **Name=my-asg**. As tags também são adicionadas a todas as instâncias executadas no grupo do Auto Scaling após essa alteração. Se uma tag com uma dessas chaves já existir, a tag existente será substituída. O console do Amazon EC2 associa o nome de exibição para cada instância ao nome especificado para a chave Name (diferencia maiúsculas de minúsculas).

```
aws autoscaling create-or-update-tags \
  --tags ResourceId=my-asg,ResourceType=auto-scaling-group,Key=Name,Value=my-
asg,PropagateAtLaunch=true \
  ResourceId=my-asg,ResourceType=auto-scaling-
group,Key=costcenter,Value=cc123,PropagateAtLaunch=true
```

Descrever as tags para um grupo do Auto Scaling (AWS CLI)

Se você deseja visualizar as tags que são aplicadas à uma função do Auto Scaling específica, pode usar os seguintes comandos:

- describe-tags: - você fornece o nome do grupo do Auto Scaling para visualizar uma lista das tags do grupo especificado.

  ```
  aws autoscaling describe-tags --filters Name=auto-scaling-group,Values=my-asg
  ```

  O seguinte é um exemplo de resposta.

  ```
  {
      "Tags": [
          {
              "ResourceType": "auto-scaling-group",
              "ResourceId": "my-asg",
              "PropagateAtLaunch": true,
              "Value": "production",
              "Key": "environment"
          }
      ]
  ```

```
}
```

- describe-auto-scaling-groups— Você fornece o nome do grupo do Auto Scaling para visualizar os atributos do grupo especificado, incluindo quaisquer tags.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

O seguinte é um exemplo de resposta.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-0b97f1e282EXAMPLE",
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "MinSize": 1,
            "MaxSize": 5,
            "DesiredCapacity": 1,
            ...
            "Tags": [
                {
                    "ResourceType": "auto-scaling-group",
                    "ResourceId": "my-asg",
                    "PropagateAtLaunch": true,
                    "Value": "production",
                    "Key": "environment"
                }
            ],
            ...
        }
    ]
}
```

Excluir tags

Você pode excluir uma tag associada a seu grupo do Auto Scaling a qualquer momento.

Conteúdo

- Excluir tags (console)
- Excluir tags (AWS CLI)

Excluir tags (console)

Para excluir uma tag

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado de um grupo existente.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Detalhes escolha Tags, Editar.
4. Escolha Remove (Remover) ao lado da tag.
5. Selecione Atualizar.

Excluir tags (AWS CLI)

Use o comando delete-tags para excluir uma tag. Por exemplo, o comando a seguir exclui uma tag com uma chave de **`environment`**.

```
aws autoscaling delete-tags --tags "ResourceId=my-asg,ResourceType=auto-scaling-
group,Key=environment"
```

É necessário especificar a chave da tag, mas você não precisa especificar o valor. Se você especificar um valor e o valor estiver incorreto, a tag não será excluída.

Etiquetas para segurança

Use etiquetas para verificar se o solicitante (como um usuário ou perfil do IAM) tem permissões para criar, modificar ou excluir grupos do Auto Scaling específicos. Forneça informações de tags no elemento de condição de uma política do IAM usando uma ou mais das seguintes chaves de condição:

- Use `autoscaling:ResourceTag/tag-key: tag-value` para permitir (ou negar) ações do usuário em grupos do Auto Scaling com tags específicas.
- Use `aws:RequestTag/tag-key: tag-value` para exigir que uma tag específica esteja presente (ou ausente) em uma solicitação.
- Use `aws:TagKeys [tag-key, ...]` para exigir que chaves de tag específicas estejam presentes (ou ausentes) em uma solicitação.

Por exemplo, você pode negar acesso a todos os grupos do Auto Scaling que incluam uma tag com a chave **`environment`** e o valor **`production`**, conforme mostrado no exemplo a seguir.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Deny",
            "Action": [
                "autoscaling:CreateAutoScalingGroup",
                "autoscaling:UpdateAutoScalingGroup",
                "autoscaling:DeleteAutoScalingGroup"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {"autoscaling:ResourceTag/environment":
 "production"}
            }
        }
    ]
}
```

Para obter mais informações sobre o uso das chaves de condição para controlar o acesso aos grupos do Auto Scaling, consulte Como o Amazon EC2 Auto Scaling funciona com o IAM.

Controlar o acesso usando etiquetas

Use etiquetas para verificar se o solicitante (como um usuário ou perfil do IAM) tem permissões para adicionar, modificar ou excluir etiquetas de grupos do Auto Scaling.

O exemplo de política do IAM a seguir fornece a permissão principal para remover apenas o tag com a chave **`temporary`** dos grupos do Auto Scaling.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "autoscaling:DeleteTags",
            "Resource": "*",
            "Condition": {
                "ForAllValues:StringEquals": { "aws:TagKeys": ["temporary"] }
            }
        }
    ]
}
```

Para ver mais exemplos de políticas do IAM que impõem restrições nas tags especificadas para grupos do Auto Scaling, consulte Controlar quais chaves de tag e valores de tag podem ser usados.

> **Note**
>
> Mesmo que você tenha uma política que restrinja os usuários de executar uma operação de marcação (ou desmarcação) em um grupo do Auto Scaling, isso não os impede de alterar manualmente as marcações nas instâncias após elas serem iniciadas. Para exemplos que controlam o acesso a tags em instâncias do EC2, consulte Exemplo: marcação de recursos no Manual do usuário do Amazon EC2 .

Usar etiquetas para filtrar grupos do Auto Scaling

Os exemplos a seguir mostram como usar filtros com o describe-auto-scaling-groupscomando para descrever grupos de Auto Scaling com tags específicas. A filtragem por tags é limitada ao AWS CLI ou a um SDK e não está disponível no console.

Considerações sobre filtragem

- É possível especificar vários filtros e vários valores de filtro em uma única solicitação.
- Não é possível usar curingas com os valores de filtro.
- Os valores do filtro diferenciam maiúsculas de minúsculas.

Exemplo: descreva grupos do Auto Scaling com um par de chave e valor de etiqueta específicos

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com a chave de etiqueta e o par de valores de **`environment=production`**.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-key,Values=environment Name=tag-value,Values=production
```

O seguinte é um exemplo de resposta.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-0b97f1e282EXAMPLE",
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "MinSize": 1,
            "MaxSize": 5,
            "DesiredCapacity": 1,
            ...
            "Tags": [
                {
                    "ResourceType": "auto-scaling-group",
                    "ResourceId": "my-asg",
                    "PropagateAtLaunch": true,
                    "Value": "production",
                    "Key": "environment"
                }
            ],
            ...
        },

    ... additional groups ...
```

```
    ]
}
```

Como alternativa, você pode especificar etiquetas usando um filtro `tag:<key>`. Por exemplo, o comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com a chave de etiqueta e o par de valores de **`environment=production`**. Este filtro é formatado da seguinte maneira: `Name=tag:<key>,Values=<value>`, com **`<key>`** e **`<value>`** representando uma etiqueta de chave e par de valor.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag:environment,Values=production
```

Você também pode filtrar a AWS CLI saída usando a `--query` opção. O exemplo a seguir mostra como limitar a AWS CLI saída do comando anterior somente ao nome do grupo, tamanho mínimo, tamanho máximo e atributos de capacidade desejados.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag:environment,Values=production \
  --query "AutoScalingGroups[].{AutoScalingGroupName: AutoScalingGroupName, MinSize:
 MinSize, MaxSize: MaxSize, DesiredCapacity: DesiredCapacity}"
```

O seguinte é um exemplo de resposta.

```
[
    {
        "AutoScalingGroupName": "my-asg",
        "MinSize": 0,
        "MaxSize": 10,
        "DesiredCapacity": 1
    },

    ... additional groups ...

]
```

Para obter mais informações sobre filtragem, consulte Filtragem AWS CLI de saída no Guia do AWS Command Line Interface usuário.

Exemplo: descreva grupos do Auto Scaling com etiquetas que correspondam à chave de etiqueta especificada

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com a etiqueta **`environment`**, independentemente do valor de etiqueta.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-key,Values=environment
```

Exemplo: descreva grupos do Auto Scaling com etiquetas que correspondam ao conjunto de chaves de etiquetas especificado

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com etiquetas para **`environment`** e **`project`**, independentemente dos valores das etiquetas.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-key,Values=environment Name=tag-key,Values=project
```

Exemplo: descreva grupos do Auto Scaling com etiquetas que correspondam a pelo menos uma das chaves de etiquetas especificadas

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com etiquetas para **`environment`** ou **`project`**, independentemente dos valores das etiquetas.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-key,Values=environment,project
```

Exemplo: descreva grupos do Auto Scaling com o valor de etiqueta especificado

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com o valor de etiqueta de **`production`**, independentemente da chave de etiqueta.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-value,Values=production
```

Exemplo: descreva grupos do Auto Scaling com o conjunto de valores de etiquetas especificado

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com os valores de **`production`** e **`development`**, independentemente da chave de etiqueta.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-value,Values=production Name=tag-value,Values=development
```

Exemplo: descreva grupos do Auto Scaling com etiquetas que correspondam a pelo menos um dos valores das etiquetas especificados

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com o valor de etiqueta de **`production`** ou **`development`**, independentemente da chave de etiqueta.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-value,Values=production,development
```

Exemplo: descreva grupos do Auto Scaling com etiquetas que correspondam a várias chaves e valores de etiquetas

Você também pode combinar filtros para criar lógicas AND e OR personalizadas para fazer uma filtragem mais complexa.

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com um conjunto específico de etiquetas. Uma chave de tag é **`environment`** AND o valor da tag é (**`production`** OR **`development`**) AND a outra chave de tag é **`costcenter`** AND o valor da tag é **`cc123`**.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag:environment,Values=production,development
 Name=tag:costcenter,Values=cc123
```

Políticas de manutenção de instância

Você pode configurar uma política de manutenção de instâncias para seu grupo do Auto Scaling para atender aos requisitos específicos de capacidade durante eventos que fazem com que as instâncias sejam substituídas, como uma atualização da instância ou o processo de verificação de integridade.

Por exemplo, suponha que você tenha um grupo do Auto Scaling com um pequeno número de instâncias. É aconselhável evitar possíveis interrupções decorrentes do encerramento e substituição de uma instância quando as verificações de integridade indicarem uma instância com defeito. Com uma política de manutenção de instâncias, você pode garantir que o Amazon EC2 Auto Scaling primeiro execute uma nova instância e depois espere que ela esteja totalmente pronta antes de encerrar a instância não íntegra.

Uma política de manutenção de instâncias também ajuda a minimizar possíveis interrupções nos casos em que várias instâncias são substituídas ao mesmo tempo. Você define os parâmetros de

porcentagem de integridade mínima e máxima para a política, e seu grupo do Auto Scaling só pode aumentar e diminuir a capacidade dentro desse intervalo mínimo-máximo ao substituir instâncias. Um intervalo maior aumenta o número de instâncias que podem ser substituídas ao mesmo tempo.

Conteúdo

- Política de manutenção de instância para o grupo do Auto Scaling
- Definir uma política de manutenção de instâncias no seu grupo do Auto Scaling

Política de manutenção de instância para o grupo do Auto Scaling

Este tópico dá uma visão geral das opções disponíveis e descreve o que deve ser considerado ao criar uma política de manutenção de instâncias.

Conteúdo

- Visão geral
- Principais conceitos
- Aquecimento da instância
- Período de carência da verificação de integridade
- Dimensionar o grupo do Auto Scaling
- Cenários de exemplo

Visão geral

Quando você cria uma política de manutenção de instâncias para seu grupo do Auto Scaling, a política afeta os eventos do Amazon EC2 Auto Scaling que fazem com que as instâncias sejam substituídas. Isso resulta em comportamentos de substituição mais consistentes dentro do mesmo grupo do Auto Scaling. Também permite otimizar seu grupo quanto à disponibilidade ou ao custo, dependendo de suas necessidades.

No console, as seguintes opções de configuração estão disponíveis:

- Iniciar antes de encerrar: uma nova instância deve ser provisionada primeiro antes que uma instância existente possa ser encerrada. Essa abordagem é uma boa opção para aplicações que favorecem a disponibilidade em detrimento da redução de custos.
- Encerrar e iniciar: novas instâncias são provisionadas ao mesmo tempo em que suas instâncias existentes são encerradas. Essa abordagem é uma boa opção para aplicações que favorecem

a redução de custos em relação à disponibilidade. Também é uma boa opção para aplicações que não devem lançar mais capacidade do que a disponível atualmente, mesmo ao substituir instâncias.

- Política personalizada – essa opção permite que você configure sua política com um intervalo mínimo e máximo personalizado para o nível de capacidade que você deseja disponibilizar ao substituir instâncias. Essa abordagem pode ajudar a alcançar o equilíbrio certo entre custo e disponibilidade.

O padrão para um grupo do Auto Scaling é não ter uma política de manutenção de instâncias, o que faz com que ele responda aos eventos de manutenção de instâncias com os comportamentos padrão. Os comportamentos padrão estão descritos na tabela a seguir.

Comportamentos padrão do evento de manutenção de instâncias

Event	Description	Comportamento padrão do
Falhas de verificação de integridade	Acontece automaticamente quando as instâncias falham nas verificações de integrida de. O Amazon EC2 Auto Scaling substitui instâncias que apresentam falhas de verificação de integridade. Para entender as causas das falhas de verificação de integridade, consulte Verificaç ões de integridade para instâncias em um grupo do Auto Scaling.	Encerrar e iniciar.
Atualização de instância	O que acontece quando você inicia uma atualização de instância. Dependendo de sua configuração, uma atualização de instância substitui instância s uma de cada vez, várias por vez ou todas de uma vez.	Encerrar e iniciar.

Event	Description	Comportamento padrão do
	Para obter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling.	
Vida útil máxima da instância	Acontece automaticamente quando as instâncias atingem a vida útil máxima que você especifica para seu grupo do Auto Scaling. O Amazon EC2 Auto Scaling substitui instâncias que atingem sua vida útil máxima. Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na vida útil máxima da instância.	Encerrar e iniciar.

Event	Description	Comportamento padrão do
Rebalanceamento	Acontece automaticamente se houver mudanças subjacentes que façam com que o grupo fique desbalanceado. O Amazon EC2 Auto Scaling reequilibra o grupo nas seguintes situações: • Uma zona de disponibilidade que antes tinha capacidade insuficiente se recupera, ou você adiciona ou remove uma zona de disponibilidade do grupo. Quando isso acontece, seu grupo do Auto Scaling tenta se equilibrar uniformemente entre as zonas de disponibilidade. Para obter mais informações, consulte Atividades de rebalanceamento. • Você ativa o rebalanceamento de capacidade em seu grupo do Auto Scaling e ele tenta iniciar novas instâncias spot antes que as existentes sejam interrompidas conforme a mudança de disponibilidade das instâncias spot. Para obter mais informações, consulte Rebalanceamento de capacidade no	Iniciar antes de encerrar. O Amazon EC2 Auto Scaling pode exceder os limites de tamanho do seu grupo em até 10% da capacidade máxima. Porém, se você estiver usando o rebalanceamento de capacidade, ele só poderá exceder esses limites em até 10% da capacidade desejada.

Event	Description	Comportamento padrão do
	Auto Scaling para substituir instâncias spot em risco. • Você atualiza seu grupo do Auto Scaling e ele substitui gradualmente as instâncias de acordo com as novas opções de compra que você escolheu ao atualizar uma política de instâncias mistas. Para obter mais informações, consulte Update an Auto Scaling group.	

O Amazon EC2 Auto Scaling continuará usando como padrão o encerramento e o lançamento nas seguintes situações. Portanto, quando uma dessas situações ocorre, a capacidade do seu grupo pode ser menor que o limite inferior da sua política de manutenção da instância.

- Quando uma instância é encerrada inesperadamente, por exemplo, devido à ação humana. O Amazon EC2 Auto Scaling substitui imediatamente instâncias que não estão mais em execução. Para obter mais informações, consulte Verificações de integridade do Amazon EC2.
- Quando o Amazon EC2 reinicia, interrompe ou desativa uma instância como parte de um evento programado antes que o Amazon EC2 Auto Scaling possa iniciar a instância substituta. Para obter mais informações sobre eventos programados, consulte Eventos programados para suas instâncias no Guia do usuário do Amazon EC2 para instâncias de Linux.
- Quando o Amazon EC2 Spot Service inicia uma interrupção de Instância spot e uma Instância Spot é encerrada à força.

Com as instâncias spot, se você habilitou o rebalanceamento de capacidade em seu grupo do Auto Scaling, talvez a instância já tenha uma instância pendente de um pool spot diferente que lançamos antes de iniciarmos a interrupção spot. Para ver mais detalhes sobre como funciona o rebalanceamento de capacidade, consulte Rebalanceamento de capacidade no Auto Scaling para substituir instâncias spot em risco.

Porém, como não é garantido que as Instâncias spot permaneçam disponíveis e possam ser encerradas com um aviso de interrupção da Instância spot de dois minutos, o limite inferior da sua política de manutenção de instâncias pode ser excedido se as instâncias forem interrompidas antes do lançamento de suas novas instâncias.

Principais conceitos

Antes de começar, familiarize-se com os seguintes conceitos e termos básicos:

Capacidade desejada

A capacidade desejada é a capacidade do grupo do Auto Scaling no momento da criação. É também a capacidade que o grupo tenta manter quando não há condições de escalabilidade associadas ao grupo.

Políticas de manutenção de instâncias

Uma política de manutenção de instâncias controla se uma instância é provisionada primeiro antes do encerramento de uma instância em eventos de manutenção de instâncias. Também determina até que ponto seu grupo do Auto Scaling pode ir abaixo e acima da capacidade desejada para substituir várias instâncias ao mesmo tempo.

Porcentagem máxima de integridade

A porcentagem máxima de integridade é a porcentagem da capacidade desejada que seu grupo do Auto Scaling pode aumentar ao substituir instâncias. Ela representa a porcentagem máxima do grupo que pode estar em serviço e íntegra, ou pendente, para suportar sua workload. No console, você pode definir a porcentagem máxima de integridade ao usar a opção Iniciar antes de encerrar ou a opção Política personalizada. Os valores válidos são 100 a 200%.

Percentual mínimo de integridade

A porcentagem mínima de integridade é a porcentagem da capacidade desejada para se manter em serviço, íntegra e pronta para ser usada a fim de suportar sua workload ao substituir instâncias. Uma instância é considerada íntegra e pronta para uso depois de concluir com êxito sua primeira verificação de integridade e após o término do tempo de aquecimento especificado. No console, você pode definir a porcentagem mínima de integridade ao usar a opção Encerrar e iniciar ou a opção Política personalizada. Os valores válidos são 0 a 100%.

> **Note**
>
> Para substituir instâncias mais rapidamente, você pode especificar uma porcentagem mínima íntegra baixa. Porém, se não houver instâncias íntegras suficientes em execução, a disponibilidade pode ser reduzida. Recomendamos selecionar um valor razoável para manter a disponibilidade em situações em que várias instâncias serão substituídas.

Aquecimento da instância

Se suas instâncias precisarem de tempo para inicializar depois de entrarem no estado`InService`, ative o aquecimento padrão da instância para seu grupo do Auto Scaling. Com o aquecimento padrão da instância, você pode evitar que as instâncias sejam contabilizadas na porcentagem mínima de integridade antes de estarem prontas. Isso garante que o Amazon EC2 Auto Scaling considere quanto tempo é necessário para ter capacidade suficiente para suportar a workload antes de encerrar as instâncias existentes.

Como benefício adicional, você pode melhorar as CloudWatch métricas da Amazon usadas para escalabilidade dinâmica ao ativar o aquecimento padrão da instância. Se seu grupo de Auto Scaling tiver alguma política de escalabilidade, quando o grupo for expandido, ele usará o mesmo período de aquecimento padrão para evitar que as instâncias sejam contabilizadas nas CloudWatch métricas antes de concluírem a inicialização.

Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Período de carência da verificação de integridade

O Amazon EC2 Auto Scaling determina se a instância está íntegra com base no status das verificações de integridade que o grupo do Auto Scaling usa. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Para garantir que essas verificações de integridade comecem o mais rápido possível, não defina um período de carência da verificação de integridade do grupo muito alto, mas alto o suficiente para que suas verificações de integridade do Elastic Load Balancing consigam determinar se um destino está disponível para lidar com solicitações. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

Dimensionar o grupo do Auto Scaling

Uma política de manutenção de instâncias só se aplica a eventos de manutenção de instâncias e não impede que o grupo seja escalado manual ou automaticamente.

Quando há políticas de escalabilidade ou ações programadas anexadas ao seu grupo do Auto Scaling, elas podem ser executadas paralelamente enquanto os eventos de manutenção da instância estão ocorrendo. Nesse caso, eles poderiam aumentar ou diminuir a capacidade desejada do grupo, mas somente dentro dos limites de escalabilidade que você definiu. Para obter mais informações sobre esses limites, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.

Cenários de exemplo

Em um cenário típico, a política de manutenção da instância e a capacidade desejada podem ser mais ou menos assim:

- Porcentagem mínima de integridade = 90%
- Porcentagem máxima de integridade = 120%
- Capacidade desejada = 100

Durante qualquer evento de manutenção de instância, seu grupo do Auto Scaling pode ter no mínimo 90 instâncias e no máximo 120. Depois do evento, o grupo volta a ter 100 instâncias.

Quando você usa uma política de manutenção de instância com um grupo do Auto Scaling que tem um pool ativo, as porcentagens de integridade mínima e máxima são aplicadas separadamente ao grupo do Auto Scaling e ao pool ativo.

Por exemplo, suponha que esta seja sua configuração:

- Porcentagem mínima de integridade = 90%
- Porcentagem máxima de integridade = 120%
- Capacidade desejada = 100
- Tamanho do pool ativo = 10

Se você iniciar uma atualização de instância para reciclar as instâncias do grupo, o Amazon EC2 Auto Scaling substituirá primeiro as instâncias no grupo do Auto Scaling e depois as instâncias

no pool ativo. Embora o Amazon EC2 Auto Scaling ainda esteja trabalhando na substituição de instâncias no grupo do Auto Scaling, o grupo pode ter no mínimo 90 instâncias e no máximo 120. Depois de terminar com o grupo, o Amazon EC2 Auto Scaling pode trabalhar na substituição de instâncias no pool ativo. Enquanto isso acontece, o pool ativo pode ter no mínimo 9 instâncias e no máximo 12.

Definir uma política de manutenção de instâncias no seu grupo do Auto Scaling

É possível criar uma política de manutenção de instâncias ao criar um grupo do Auto Scaling. Também é possível criá-la para grupos existentes.

Ao definir uma política de manutenção de instância no seu grupo do Auto Scaling, não é mais necessário especificar valores de parâmetros de porcentagem mínima e máxima de integridade para o recurso de atualização da instância a não ser que queira substituir a política de manutenção de instâncias.

No console, o Amazon EC2 Auto Scaling fornece opções para ajudar você a começar.

Conteúdo

- Definir uma política de manutenção de instâncias
- Remover uma política de manutenção de instância

Definir uma política de manutenção de instâncias

Para definir uma política de manutenção de instâncias em um grupo do Auto Scaling, use um dos seguintes métodos:

Console

Para definir uma política de manutenção de instâncias em um novo grupo (console)

1. Siga as instruções em Criar um grupo do Auto Scaling usando um modelo de execução e conclua cada etapa do procedimento, até a etapa 11.
2. Em Configurar tamanho do grupo e políticas de escalabilidade, em Capacidade desejada, insira o número inicial de instâncias a serem executadas.
3. Na seção Escalabilidade, em Limites de escalabilidade, se o novo valor para Capacidade desejada for maior que Capacidade mínima desejada e Capacidade máxima desejada,

a Capacidade máxima desejada será automaticamente aumentada para o novo valor da capacidade desejada. Você pode alterar esses limites conforme necessário.

4. Em Escalabilidade automática, escolha se você deseja criar uma política de escalabilidade de rastreamento de destino. Você também pode criar essa política depois de criar seu grupo do Auto Scaling.

   Se você escolher a política de escalabilidade de rastreamento de destino, siga as instruções em Criar uma política de dimensionamento com monitoramento do objetivo para criar a política.

5. Na seção Política de manutenção de instâncias, escolha uma das opções disponíveis:

   - Iniciar antes de encerrar: uma nova instância deve ser provisionada primeiro antes que uma instância existente possa ser encerrada. Essa é uma boa opção para aplicações que favorecem a disponibilidade em detrimento da redução de custos.
   - Encerrar e iniciar: novas instâncias são provisionadas ao mesmo tempo em que as instâncias existentes são encerradas. Esta é uma boa opção para aplicações que favorecem a economia de custos em detrimento da disponibilidade. Também é uma boa opção para aplicações que não devem lançar mais capacidade do que a disponível atualmente.
   - Política personalizada: essa opção permite que você configure sua política com um intervalo mínimo e máximo personalizado para o nível de capacidade que você deseja disponibilizar ao substituir instâncias. Isso pode ajudar a alcançar o equilíbrio certo entre custo e disponibilidade.

6. Em Definir porcentagem de integridade, insira valores para um ou ambos os campos a seguir. Os campos habilitados variam de acordo com a opção escolhida na etapa anterior.

   - Mínimo: define a porcentagem mínima de integridade necessária para continuar com a substituição de instâncias.
   - Máximo: define a porcentagem máxima de integridade possível ao substituir instâncias.

7. Expanda a seção Exibir capacidade durante as substituições com base na seção de capacidade desejada para confirmar como os valores de Mínimo e Máximo são aplicados ao seu grupo. Os valores exatos usados dependem do valor de capacidade desejado, que mudará se o grupo for ampliado.

8. Continue com as etapas em Criar um grupo do Auto Scaling usando um modelo de execução.

AWS CLI

Para definir uma política de manutenção de instância em um novo grupo (AWS CLI)

Adicione a `--instance-maintenance-policy` opção ao create-auto-scaling-groupcomando. O exemplo a seguir define uma política de manutenção de instâncias em um novo grupo do Auto Scaling chamado *my-asg*.

```
aws autoscaling create-auto-scaling-group \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --auto-scaling-group-name my-asg \
  --min-size 1 \
  --max-size 10 \
  --desired-capacity 5 \
  --default-instance-warmup 20 \
  --instance-maintenance-policy '{
      "MinHealthyPercentage": 90,
      "MaxHealthyPercentage": 120
    }' \
  --vpc-zone-identifier "subnet-5e6example,subnet-613example,subnet-c93example"
```

Console

Para definir uma política de manutenção de instância em um grupo existente (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a mesma Região da AWS na qual você criou o grupo do Auto Scaling.
3. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
4. Na guia Detalhes, escolha Política de manutenção de instâncias, Editar.
5. Para definir uma política de manutenção de instância no grupo, escolha uma das opções disponíveis:

   - Iniciar antes de encerrar: uma nova instância deve ser provisionada primeiro antes que uma instância existente possa ser encerrada. Essa é uma boa opção para aplicações que favorecem a disponibilidade em detrimento da redução de custos.

- Encerrar e iniciar: novas instâncias são provisionadas ao mesmo tempo em que as instâncias existentes são encerradas. Esta é uma boa opção para aplicações que favorecem a economia de custos em detrimento da disponibilidade. Também é uma boa opção para aplicações que não devem lançar mais capacidade do que a disponível atualmente.
- Política personalizada: essa opção permite que você configure sua política com um intervalo mínimo e máximo personalizado para o nível de capacidade que você deseja disponibilizar ao substituir instâncias. Isso pode ajudar a alcançar o equilíbrio certo entre custo e disponibilidade.

6. Em Definir porcentagem de integridade, insira valores para um ou ambos os campos a seguir. Os campos habilitados variam de acordo com a opção escolhida na etapa anterior.

   - Mínimo: define a porcentagem mínima de integridade necessária para continuar com a substituição de instâncias.
   - Máximo: define a porcentagem máxima de integridade possível ao substituir instâncias.

7. Expanda a seção Exibir capacidade durante as substituições com base na seção de capacidade desejada para confirmar como os valores de Mínimo e Máximo são aplicados ao seu grupo. Os valores exatos usados dependem do valor de capacidade desejado, que mudará se o grupo for ampliado.
8. Selecione Atualizar.

AWS CLI

Para definir uma política de manutenção de instância em um grupo existente (AWS CLI)

Adicione a `--instance-maintenance-policy` opção ao update-auto-scaling-groupcomando. O exemplo a seguir define uma política de manutenção de instância no grupo do Auto Scaling especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --instance-maintenance-policy '{
      "MinHealthyPercentage": 90,
      "MaxHealthyPercentage": 120
    }'
```

Remover uma política de manutenção de instância

Se você quiser parar de usar uma política de manutenção de instâncias com seu grupo do Auto Scaling, você pode removê-la.

Console

Para remover uma política de manutenção de instância (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a mesma Região da AWS na qual você criou o grupo do Auto Scaling.
3. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
4. Na guia Detalhes, escolha Política de manutenção de instâncias, Editar.
5. Escolha Nenhuma política de manutenção de instâncias.
6. Escolha Atualizar.

AWS CLI

Para remover uma política de manutenção de instâncias (AWS CLI)

Adicione a `--instance-maintenance-policy` opção ao update-auto-scaling-groupcomando. O exemplo a seguir remove a política de manutenção de instâncias do grupo do Auto Scaling especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --instance-maintenance-policy '{
      "MinHealthyPercentage": -1,
      "MaxHealthyPercentage": -1
    }'
```

Ganchos do ciclo de vida do Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling oferece a capacidade de adicionar ganchos do ciclo de vida aos seus grupos do Auto Scaling. Esses hooks permitem criar soluções que estejam ciente de eventos no ciclo

de vida da instância do Auto Scaling e, em seguida, executar uma ação personalizada em instâncias quando ocorrer o evento do ciclo de vida correspondente. Um gancho do ciclo de vida fornece uma quantidade especificada de tempo (uma hora, por padrão) para esperar a ação completar antes que a instância faça a transição para o próximo estado.

Como exemplo do uso de ganchos do ciclo de vida com instâncias do Auto Scaling:

- Quando ocorre um evento de aumento da escala na horizontal, sua instância recém-iniciada conclui a sequência de inicialização e faz a transição para um estado de espera. Enquanto a instância está em um estado de espera, ela executa um script para baixar e instalar os pacotes de software necessários para sua aplicação, garantindo que sua instância esteja totalmente pronta antes de começar a receber tráfego. Quando o script terminar de instalar o software, ele envia o comando complete-lifecycle-action para continuar.
- Quando ocorre um evento de escalabilidade, um gancho de ciclo de vida pausa a instância antes que ela seja encerrada e envia uma notificação usando a Amazon. EventBridge Enquanto a instância estiver em estado de espera, você pode invocar uma AWS Lambda função ou conectar-se à instância para baixar registros ou outros dados antes que a instância seja totalmente encerrada.

Um uso popular de ganchos do ciclo de vida é controlar quando as instâncias são registradas com o Elastic Load Balancing. Ao adicionar um gancho do ciclo de vida de execução ao seu grupo do Auto Scaling, você pode garantir que seus scripts de bootstrap foram completados com êxito e que as aplicações nas instâncias estejam prontas para aceitar tráfego antes de serem registradas no balanceador de carga no final do gancho do ciclo de vida.

Conteúdo

- Disponibilidade de ganchos do ciclo de vida
- Considerações e limitações dos ganchos do ciclo de vida
- Recursos relacionados
- Como os ganchos do ciclo de vida funcionam em grupos do Auto Scaling
- Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling
- Controle a retenção de instâncias com políticas de ciclo de vida de instâncias
- Recuperar o estado de destino do ciclo de vida por meio de metadados de instância
- Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling
- Concluir uma ação do ciclo de vida em um grupo do Auto Scaling

- Tutorial: Usar script de dados e metadados de instância para recuperar o estado do ciclo de vida
- Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda

Disponibilidade de ganchos do ciclo de vida

A tabela a seguir lista os ganchos do ciclo de vida disponíveis para vários cenários.

Event	Início ou término da instância[1]	Maximum Instance Lifetime (Tempo de vida máximo da instância): instâncias de substituição	Instance Refresh (Atualizações de instância): instâncias de substituição	Capacity Rebalancing (Rebalanceamento de capacidade): instâncias de substituição	Warm Pools (Grupos de alta atividade): instâncias entrando e saindo do grupo de alta atividade
Início de instâncias	✓	✓	✓	✓	✓
Término de instâncias	✓	✓	✓	✓	✓

[1] Aplica-se a todos as execuções e encerramentos, sejam eles iniciados automática ou manualmente, por exemplo, quando você chama as operações `SetDesiredCapacity` ou `TerminateInstanceInAutoScalingGroup`. Não se aplica quando você anexa ou desvincula instâncias, move instâncias dentro e fora do modo de espera ou exclui o grupo com a opção force delete (forçar exclusão).

Considerações e limitações dos ganchos do ciclo de vida

Ao trabalhar com ganchos do ciclo de vida, tenha em mente as seguintes notas e limitações:

- O Amazon EC2 Auto Scaling fornece seu próprio ciclo de vida para ajudar no gerenciamento de grupos do Auto Scaling. Esse ciclo de vida é diferente do de outras instâncias do EC2. Para obter mais informações, consulte Ciclo de vida das instâncias do Amazon EC2 Auto Scaling. As instâncias em um grupo de alta atividade também têm seu próprio ciclo de vida, conforme descrito em Transições de estado do ciclo de vida para instâncias em um pool ativo.

- Por padrão, os ganchos do ciclo de vida da terminação operam com base no melhor esforço. Se um ciclo de vida de encerramento expirar ou for abandonado, o Amazon EC2 Auto Scaling prosseguirá com o encerramento imediato da instância. Você pode combinar ganchos de ciclo de vida de encerramento com uma política de ciclo de vida de instâncias para retenção de instâncias. Para obter mais informações, consulte Controle a retenção de instâncias com políticas de ciclo de vida de instâncias.
- Você pode usar ganchos do ciclo de vida com instâncias spot, mas um gancho do ciclo de vida não impede que uma instância seja terminada em caso de a capacidade não estar mais disponível, o que pode acontecer a qualquer momento, com um aviso de interrupção de dois minutos. Para obter mais informações, consulte Interrupções da instância spot no Guia do usuário do Amazon EC2. No entanto, você pode habilitar o rebalanceamento de capacidade para substituir proativamente as instâncias spot que receberam uma recomendação de rebalanceamento do Amazon EC2 Spot Service, um sinal que é enviado quando uma instância spot está em risco elevado de interrupção. Para obter mais informações, consulte Rebalanceamento de capacidade no Auto Scaling para substituir instâncias spot em risco.
- As instâncias podem permanecer em um estado de espera por um determinado período de tempo. O tempo limite padrão para um gancho do ciclo de vida é de uma hora (tempo limite de pulsação). Também há um tempo limite global que especifica a quantidade máxima de tempo que você pode manter uma instância em um estado de espera. O tempo limite global é de 48 horas ou 100 vezes o tempo limite de pulsação, o que for mais curto.
- O resultado do gancho do ciclo de vida pode ser abandonar ou continuar. Se uma instância estiver sendo executada, continue indica que suas ações foram bem-sucedidas e que o Amazon EC2 Auto Scaling pode colocar a instância em serviço. Caso contrário, abandonar indica que suas ações personalizadas não tiveram êxito e que podemos encerrar e substituir a instância. Se uma instância estiver sendo encerrada, abandone e continue permita que a instância seja encerrada. No entanto, abandon (abandonar) interrompe quaisquer ações restantes, como outros ganchos do ciclo de vida, e continue (continuar) permite que quaisquer outros ganchos do ciclo de vida sejam concluídos.
- O Amazon EC2 Auto Scaling limita a taxa na qual permite que as instâncias sejam iniciadas se os ganchos do ciclo de vida estiverem falhando de maneira consistente. Portanto, verifique e corrija erros permanentes em suas ações do ciclo de vida.
- Criar e atualizar ganchos de ciclo de vida usando o AWS CLI, CloudFormation, ou um SDK fornece opções não disponíveis ao criar um gancho de ciclo de vida a partir do. Console de gerenciamento da AWS Por exemplo, o campo para especificar o ARN de um tópico do SNS ou fila do SQS não aparece no console, porque o Amazon EC2 Auto Scaling já envia eventos para a Amazon.

EventBridge Esses eventos podem ser filtrados e redirecionados para AWS serviços como Lambda, Amazon SNS e Amazon SQS, conforme necessário.

- Você pode adicionar vários ganchos de ciclo de vida a um grupo do Auto Scaling enquanto o cria, chamando a CreateAutoScalingGroupAPI usando o AWS CLI,, ou um SDK. CloudFormation No entanto, cada hook deve ter o mesmo destino de notificação e perfil do IAM, se esses elementos forem especificados. Para criar ganchos de ciclo de vida com diferentes alvos de notificação e funções diferentes, crie os ganchos de ciclo de vida um por vez em chamadas separadas para a API. PutLifecycleHook
- Se você adicionar um gancho do ciclo de vida para a execução da instância, o período de carência da verificação de integridade será iniciado assim que a instância atingir o estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

Considerações sobre dimensionamento

- As políticas de escalabilidade dinâmica aumentam e diminuem em resposta a dados CloudWatch métricos, como CPU e E/S de rede, que são agregados em várias instâncias. Quando há uma expansão, o Amazon EC2 Auto Scaling não conta imediatamente uma nova instância para as métricas agregadas de instância do grupo do Auto Scaling. Ele espera até que a instância atinja o estado `InService` e o aquecimento da instância seja concluído. Para obter mais informações, consulte Considerações sobre o desempenho de escalabilidade o tópico de aquecimento da instância padrão.
- Na redução da escala na horizontal, talvez as métricas agregadas da instância não reflitam instantaneamente a remoção de uma instância de encerramento. A instância de encerramento para de contabilizar as métricas agregadas de instância do grupo pouco após o início do fluxo de trabalho de encerramento do Amazon EC2 Auto Scaling.
- Na maioria dos casos, quando os ganchos do ciclo de vida são invocados, as atividades de escalabilidade devido a políticas de escalabilidade simples são pausadas até que as ações do ciclo de vida sejam concluídas e o período de esfriamento expire. A definição de um intervalo longo para o período de desaquecimento significa que a retomada da escalabilidade levará mais tempo. Para obter mais informações, consulte Ganchos do ciclo de vida podem causar mais atrasos o tópico de resfriamento. Em geral, não recomendamos o uso de políticas de escalabilidade simples se você puder usar políticas de escalabilidade por etapas ou rastreamento de metas.

Recursos relacionados

Para ver um vídeo de introdução, consulte AWS re:Invent 2018: Gerenciamento de capacidade facilitado com o Amazon EC2 Auto Scaling ativado. YouTube

Fornecemos alguns trechos de modelos JSON e YAML que você pode usar para entender como declarar ganchos de ciclo de vida em seus modelos de pilha. CloudFormation Para obter mais informações, consulte a AWS::AutoScaling::LifecycleHookreferência no Guia AWS CloudFormation do usuário.

Você também pode visitar nosso GitHubrepositório para baixar exemplos de modelos e scripts de dados do usuário para ganchos de ciclo de vida.

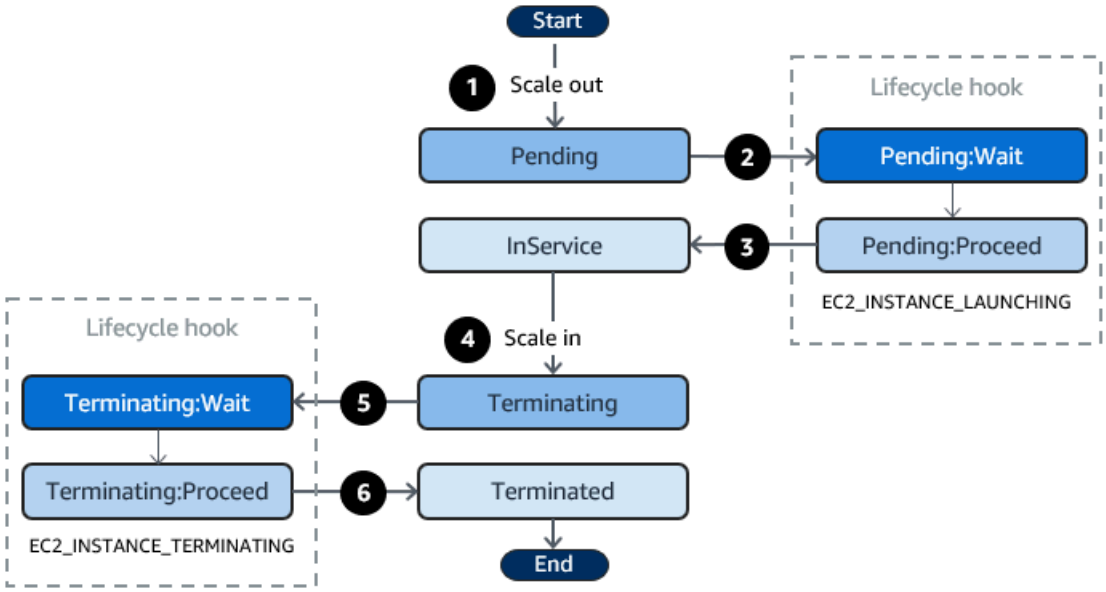
Para obter exemplos do uso de ganchos do ciclo de vida, consulte as seguintes postagens no blog.

- Construir um sistema de backup para instâncias escaláveis usando o comando de execução Lambda e Amazon EC2
- Execute o código antes de encerrar uma instância do EC2 Auto Scaling.

Como os ganchos do ciclo de vida funcionam em grupos do Auto Scaling

Uma instância do Amazon EC2 passa por diferentes estados do momento em que é iniciada até seu término. Você pode criar ações personalizadas para que seu grupo do Auto Scaling atue quando uma instância transitar para um estado de espera devido a um gancho do ciclo de vida.

A ilustração a seguir mostra as transições entre os estados de instância do Auto Scaling quando você usa ganchos do ciclo de vida para aumentar e reduzir a escala horizontalmente.



Conforme mostrado no diagrama anterior:

1. O grupo do Auto Scaling responde a um evento de aumento de escala na horizontal e começa a iniciar uma instância.
2. O gancho do ciclo de vida coloca a instância em um estado de espera (`Pending:Wait`) e, em seguida, executa uma ação personalizada.

   A instância permanece em um estado de espera até que você conclua a ação do ciclo de vida ou até o período de tempo limite terminar. Por padrão, a instância permanece em estado de espera por uma hora e, em seguida, o grupo do Auto Scaling continua o processo de início (`Pending:Proceed`). Se precisar de mais tempo, você poderá reiniciar o período de tempo limite registrando uma pulsação. Se você concluir a ação do ciclo de vida quando a ação personalizada estiver concluída e o período de tempo limite ainda não tiver expirado, o período terminará e o grupo do Auto Scaling continuará o processo de execução.
3. A instância entra no estado `InService` e o período de carência da verificação de integridade é iniciado. Contudo, antes da instância atingir o estado `InService`, se o grupo do Auto Scaling estiver associado a um balanceador de carga Elastic Load Balancing, a instância será registrada no balanceador de carga e o balanceador de carga começará a verificar sua integridade. Após o término do período de carência da verificação de integridade, o Amazon EC2 Auto Scaling começa a verificar o estado de integridade da instância.
4. O grupo do Auto Scaling responde a um evento de redução de escala na horizontal e começa a terminar uma instância. Se o grupo do Auto Scaling estiver sendo usado com o Elastic Load Balancing, primeiro é cancelado o registro da instância em término no balanceador de carga. Se

a descarga da conexão estiver habilitada para o balanceador de carga, a instância deixará de aceitar novas conexões e aguardará até que as conexões existentes sejam descarregadas antes de concluir o processo de cancelamento do registro.

5. O gancho do ciclo de vida coloca a instância em um estado de espera (`Terminating:Wait`) e, em seguida, executa uma ação personalizada.

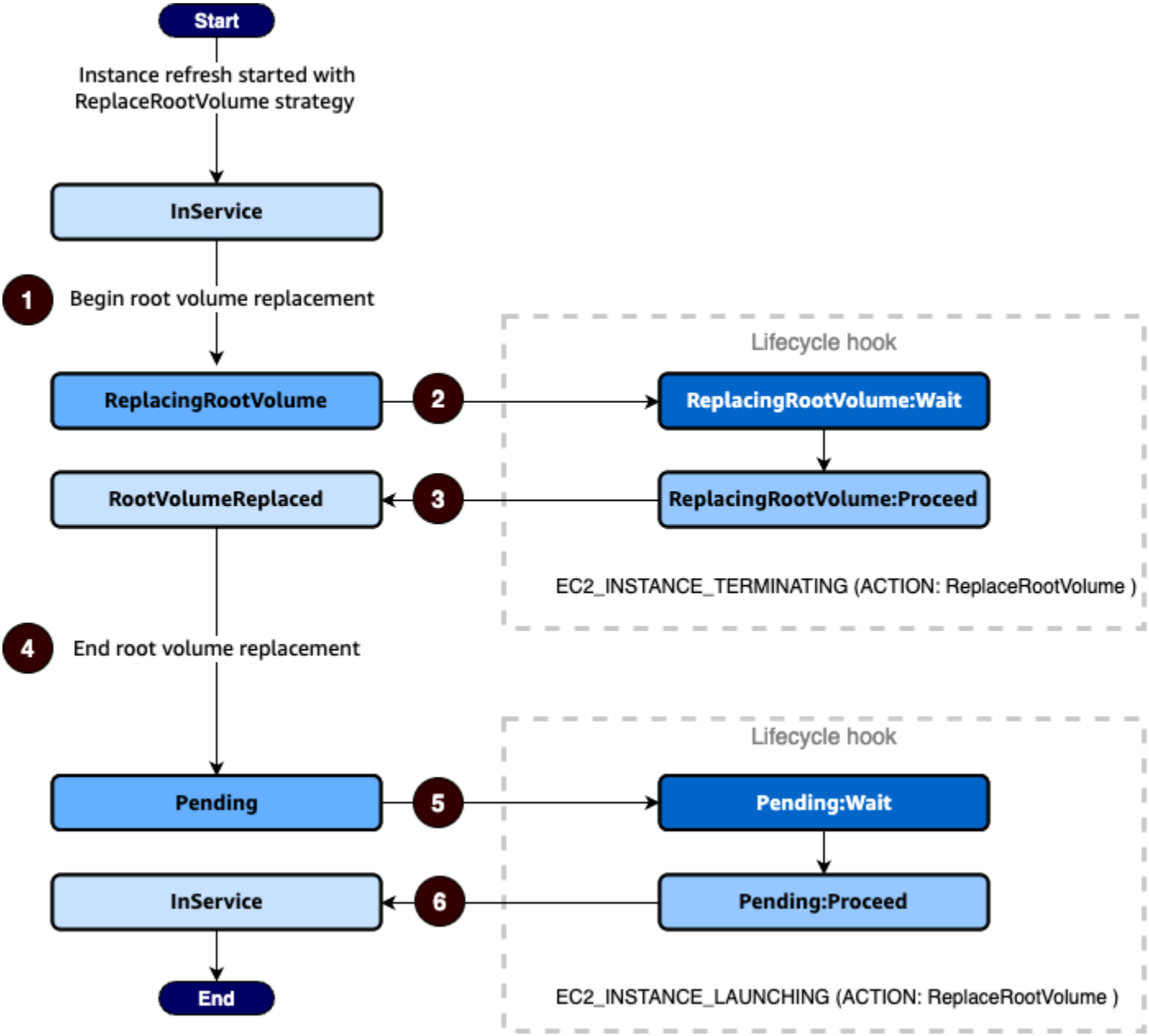
   A instância permanece em um estado de espera até que você conclua a ação do ciclo de vida, ou até o período de tempo limite terminar (uma hora, por padrão). Depois de concluir o gancho do ciclo de vida ou do período de tempo limite expirar, a instância passa para o próximo estado (`Terminating:Proceed`).

6. A instância está terminada.

> **Important**
>
> As instâncias em um pool ativo também têm seu próprio ciclo de vida com estados de espera correspondentes, conforme descrito em Transições de estado do ciclo de vida para instâncias em um pool ativo.

Transições de estado do ciclo de vida para instâncias que estão passando por substituição do volume raiz

O diagrama a seguir mostra a transição entre os estados da instância do Auto Scaling quando você usa ganchos de ciclo de vida para substituir o volume raiz:



Conforme mostrado no diagrama anterior:

1. O grupo Auto Scaling responde a uma atualização de instância e seleciona uma instância para substituição do volume raiz. A instância entra no `ReplacingRootVolume` estado. Se a instância estiver registrada com um balanceador de carga, seu registro será cancelado do balanceador de carga.
2. O gancho do ciclo de vida coloca a instância em um estado de espera (`ReplacingRootVolume:Wait`) e, em seguida, executa uma ação personalizada. A instância permanece em um estado de espera até que você conclua a ação do ciclo de vida ou até o

período de tempo limite terminar. Se você concluir a ação do ciclo de vida quando a ação personalizada for concluída e o período de tempo limite ainda não tiver expirado, o período terminará e o grupo Auto Scaling continuará o processo de substituição do volume raiz.

3. A instância conclui a substituição do volume raiz e entra no `RootVolumeReplaced` estado.
4. A instância entra no `Pending` estado.
5. O gancho do ciclo de vida coloca a instância em um estado de espera (`Pending:Wait`) e, em seguida, executa uma ação personalizada. A instância permanece em um estado de espera até que você conclua a ação do ciclo de vida ou até que o período de tempo limite termine. Depois de concluir o gancho do ciclo de vida ou do período de tempo limite expirar, a instância passa para o próximo estado (`Pending:Proceed`).
6. A instância entra no `InService` estado. No entanto, antes que a instância atinja o `InService` estado, se o grupo Auto Scaling estiver associado a um load balancer do Elastic Load Balancing, a instância será registrada no balanceador de carga.

Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling

Antes de adicionar um gancho do ciclo de vida ao grupo do Auto Scaling, certifique-se de que o script de dados do usuário ou o destino de notificação esteja configurado corretamente.

- Não é necessário configurar um destino de notificação para usar um script de dados do usuário a fim de executar ações personalizadas em suas instâncias enquanto elas estão sendo iniciadas. No entanto, você já deverá ter criado o modelo de execução ou a configuração de execução que especifica o script de dados do usuário e associado ao seu grupo do Auto Scaling. Para obter mais informações sobre scripts de dados do usuário, consulte Executar comandos na instância do Linux na inicialização no Guia do usuário do Amazon EC2 para instâncias do Linux.
- Para sinalizar ao Amazon EC2 Auto Scaling quando a ação do ciclo de vida estiver concluída, você deve CompleteLifecycleActionadicionar a chamada da API ao script e criar manualmente uma função do IAM com uma política que permita que as instâncias do Auto Scaling chamem essa API. Seu modelo de execução ou configuração de execução deve especificar essa função usando um perfil de instância do IAM que é anexado às suas instâncias do Amazon EC2 na inicialização. Para obter mais informações, consulte Concluir uma ação do ciclo de vida em um grupo do Auto Scaling e Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2.
- Para permitir que o Lambda sinalize ao Amazon EC2 Auto Scaling quando a ação do ciclo de vida for concluída, você deve adicionar CompleteLifecycleActiona chamada de API ao código da

função. Você também deve ter anexado uma política do IAM à função de execução da função que concede permissão ao Lambda para concluir ações do ciclo de vida. Para obter mais informações, consulte Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda.

- Para usar um serviço como um Amazon SNS ou Amazon SQS para executar uma ação personalizada, você já deverá ter criado o tópico do SNS ou a fila do SQS e ter pronto o nome do recurso da Amazon (ARN). Você também já deverá ter criado o perfil do IAM que concede ao Amazon EC2 Auto Scaling acesso ao tópico do SNS ou ao destino do SQS e ter pronto o ARN. Para obter mais informações, consulte Configurar um destino de notificação para notificações do ciclo de vida.

> **Note**
>
> Por padrão, quando você adiciona um gancho de ciclo de vida no console, o Amazon EC2 Auto Scaling envia notificações de eventos de ciclo de vida para a Amazon. EventBridge Usar EventBridge ou usar um script de dados do usuário é uma prática recomendada. Para criar um gancho de ciclo de vida que envie notificações diretamente para o Amazon SNS, Amazon SQS, ou use o, AWS CloudFormation, AWS Lambda ou um SDK para adicionar AWS CLI o gancho de ciclo de vida.

Configurar um destino de notificação para notificações do ciclo de vida

Você pode adicionar ganchos do ciclo de vida a um grupo do Auto Scaling para executar ações personalizadas sempre que uma instância entrar em um estado de espera. Você pode escolher um serviço de destino para executar essas ações dependendo de sua abordagem de desenvolvimento preferida.

Há quatro abordagens diferentes para implementar destinos de notificação para ganchos do ciclo de vida:

- Amazon EventBridge — Receba as notificações e execute as ações que você deseja.
- Amazon Simple Notification Service (Amazon SNS): cria um tópico para publicar notificações. Os clientes podem se inscrever no tópico do SNS e receber mensagens publicadas usando um protocolo compatível.
- Amazon Simple Queue Service (Amazon SQS): troca mensagens por meio de um modelo de pesquisa.
- AWS Lambda: invoca uma função do Lambda que executa a ação desejada.

Como prática recomendada, sugerimos que você use EventBridge. As notificações enviadas para o Amazon SNS e o Amazon SQS contêm as mesmas informações que as notificações para as quais o Amazon EC2 Auto Scaling envia. EventBridge Antes EventBridge, a prática padrão era enviar uma notificação ao SNS ou ao SQS e integrar outro serviço ao SNS ou SQS para realizar ações programáticas. Hoje, EventBridge oferece mais opções para quais serviços você pode segmentar e facilita o tratamento de eventos usando a arquitetura sem servidor.

Lembre-se de que, se você tiver um script de dados do usuário no modelo de execução ou uma configuração de execução que configure suas instâncias quando forem iniciadas, você não precisa receber notificações para executar ações personalizadas em suas instâncias.

Os procedimentos a seguir abordam como configurar seu destino de notificação.

Conteúdo

- Encaminhe notificações para o Lambda usando EventBridge
- Receba notificações usando o Amazon SNS
- Receba notificações usando o Amazon SQS
- Encaminhe notificações AWS Lambda diretamente para
- Exemplo de mensagem de notificação

> **Important**
>
> A EventBridge regra, a função Lambda, o tópico do Amazon SNS e a fila do Amazon SQS que você usa com ganchos de ciclo de vida devem estar sempre na mesma região em que você criou seu grupo de Auto Scaling.

Encaminhe notificações para o Lambda usando EventBridge

Você pode configurar uma EventBridge regra para invocar uma função Lambda quando uma instância entra em um estado de espera. O Amazon EC2 Auto Scaling emite uma EventBridge notificação de evento do ciclo de vida sobre a instância que está sendo iniciada ou encerrada e um token que você pode usar para controlar a ação do ciclo de vida. Para obter exemplos desses eventos, consulte Referência de eventos do Amazon EC2 Auto Scaling.

> **Note**
>
> Quando você usa o Console de gerenciamento da AWS para criar uma regra de evento, o console adiciona automaticamente as permissões do IAM necessárias para conceder EventBridge permissão para chamar sua função Lambda. Caso esteja criando uma regra de evento usando a AWS CLI, você precisa conceder essa permissão explicitamente.
> Para obter informações sobre como criar regras de eventos no EventBridge console, consulte Criação de EventBridge regras da Amazon que reagem a eventos no Guia EventBridge do usuário da Amazon.
> - ou -
> Para um tutorial introdutório direcionado a usuários do console, consulte Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda. Este tutorial mostra como criar uma função Lambda simples que escuta os eventos de lançamento e os grava em um CloudWatch registro de registros.

Para criar uma EventBridge regra que invoque uma função Lambda

1. Crie uma função do Lambda usando o console do Lambda e observe seu nome de recurso da Amazon (ARN). Por exemplo, `.arn:aws:lambda:region:123456789012:function:my-function` Você precisa do ARN para criar um EventBridge destino. Para obter mais informações, consulte Conceitos básicos do Lambda no Guia do desenvolvedor do AWS Lambda .
2. Para criar uma regra que faça a correspondência com eventos para inicialização de instâncias, use o seguinte comando put-rule.

   ```
   aws events put-rule --name my-rule --event-pattern file://pattern.json --state
    ENABLED
   ```

   O exemplo a seguir mostra o `pattern.json` de uma ação do ciclo de vida de execício ução de instância. Substitua o texto pelo nome do seu grupo de Auto Scaling. *italics*

   ```
   {
     "source": [ "aws.autoscaling" ],
     "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
     "detail": {
         "AutoScalingGroupName": [ "my-asg" ]
      }
   ```

```
}
```

Se o comando for executado com êxito, EventBridge responderá com o ARN da regra. Anote esse ARN. Você vai precisar dele na etapa 4.

Para criar uma regra que faça a correspondência com outros eventos, modifique o padrão de evento. Para obter mais informações, consulte Use EventBridge para lidar com eventos do Auto Scaling.

3. Para especificar a função do Lambda a ser usada como destino para a regra, use o seguinte comando put-targets.

```
aws events put-targets --rule my-rule --targets
 Id=1,Arn=arn:aws:lambda:region:123456789012:function:my-function
```

No comando anterior, *my-rule* é o nome que você especificou para a regra na etapa 2 e o valor do `Arn` parâmetro é o ARN da função que você criou na etapa 1.

4. Para adicionar permissões que deixem a regra invocar sua função do Lambda, use o seguinte comando add-permission do Lambda. Esse comando confia no EventBridge service principal (`events.amazonaws.com`) e define o escopo das permissões para a regra especificada.

```
aws lambda add-permission --function-name my-function --statement-id my-unique-id \
  --action 'lambda:InvokeFunction' --principal events.amazonaws.com --source-arn
 arn:aws:events:region:123456789012:rule/my-rule
```

No comando anterior:

- *my-function*é o nome da função Lambda que você deseja que a regra use como destino.
- *my-unique-id*é um identificador exclusivo que você define para descrever a declaração na política da função Lambda.
- `source-arn`é o ARN da regra. EventBridge

Se o comando for executado com êxito, você receberá um resultado semelhante a este.

```
{
  "Statement": "{\"Sid\":\"my-unique-id\",
    \"Effect\":\"Allow\",
    \"Principal\":{\"Service\":\"events.amazonaws.com\"},
```

```
    \"Action\":\"lambda:InvokeFunction\",
    \"Resource\":\"arn:aws:lambda:us-west-2:123456789012:function:my-function\",
    \"Condition\":
      {\"ArnLike\":
        {\"AWS:SourceArn\":
         \"arn:aws:events:us-west-2:123456789012:rule/my-rule\"}}}"
}
```

O valor de `Statement` é uma versão da string JSON da instrução adicionada à política da função do Lambda.

5. Depois que você tiver seguido estas instruções, prossiga para Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling como próxima etapa.

Receba notificações usando o Amazon SNS

Você pode usar o Amazon SNS para configurar um destino de notificação (um tópico do SNS) para receber notificações quando ocorrer uma ação do ciclo de vida. Em seguida, o Amazon SNS envia as notificações para os destinatários inscritos. Até que a inscrição seja confirmada, nenhuma notificação publicada no tópico é enviada para os destinatários.

Para configurar notificações usando o Amazon SNS

1. Crie um tópico do Amazon SNS usando o console do Amazon SNS ou o seguinte comando create-topic. Verifique se o tópico está na mesma região do grupo do Auto Scaling que você está usando. Para obter mais informações, consulte Conceitos básicos do Amazon SNS no Guia do desenvolvedor do Amazon Simple Notification Service.

```
aws sns create-topic --name my-sns-topic
```

2. Observe o nome de recurso da Amazon (ARN) do tópico, por exemplo, `arn:aws:sns:region:123456789012:my-sns-topic`. Você precisa dele para criar o gancho do ciclo de vida.
3. Crie uma função de serviço do IAM para dar ao Amazon EC2 Auto Scaling acesso ao seu destino de notificação do Amazon SNS.

   Para dar ao Amazon EC2 Auto Scaling acesso ao seu tópico do SNS

   a. Abra o console do IAM em https://console.aws.amazon.com/iam/.
   b. No painel de navegação à esquerda, escolha Roles (Funções).

c. Selecione Criar perfil.

d. Em Select trusted entity (Selecionar entidade confiável), escolha AWS service (serviço).

e. Para seu caso de uso, em Use cases for other AWS services (Casos de uso de outros produtos), escolha EC2 Auto Scaling e depois EC2 Auto Scaling Notification Access (Acesso à notificação do EC2 Auto Scaling).

f. Escolha Next (Próximo) duas vezes para ir até a página Name, review, and create (Nomear, revisar e criar).

g. Em Role name (Nome da função), insira um nome para a função (por exemplo, **`my-notification-role`**) e escolha Create role (Criar função).

h. Na página Roles (Funções), escolha a função recém-criada para abrir a página Summary (Resumo). Anote o Role ARN (ARN da função). Por exemplo, `.arn:aws:iam::123456789012:role/my-notification-role` Você precisa dele para criar o gancho do ciclo de vida.

4. Depois que você tiver seguido estas instruções, prossiga para Adicionar ganchos do ciclo de vida (AWS CLI) como próxima etapa.

Receba notificações usando o Amazon SQS

Você pode usar o Amazon SQS para configurar um destino de notificação para receber notificações quando ocorrer uma ação do ciclo de vida. Um consumidor da fila deve sondar uma fila do SQS para agir nessas notificações.

> **Important**
>
> As filas FIFO não são compatíveis com ganchos do ciclo de vida.

Para configurar notificações usando o Amazon SQS

1. Crie uma fila do Amazon SQS usando o console do Amazon SQS. Verifique se a fila está na mesma região do grupo do Auto Scaling que você está usando. Para obter mais informações, consulte Conceitos básicos do Amazon SQS no Guia do desenvolvedor do Amazon Simple Queue Service.
2. Observe o ARN da fila, por exemplo, `arn:aws:sqs:us-west-2:123456789012:my-sqs-queue`. Você precisa dele para criar o gancho do ciclo de vida.

3. Crie uma função de serviço do IAM para dar ao Amazon EC2 Auto Scaling acesso ao seu destino de notificação do Amazon SQS.

   Para dar ao Amazon EC2 Auto Scaling acesso à sua fila do SQS

   a. Abra o console do IAM em https://console.aws.amazon.com/iam/.
   b. No painel de navegação à esquerda, escolha Roles (Funções).
   c. Selecione Criar perfil.
   d. Em Select trusted entity (Selecionar entidade confiável), escolha AWS service (serviço).
   e. Para seu caso de uso, em Use cases for other AWS services (Casos de uso de outros produtos), escolha EC2 Auto Scaling e depois EC2 Auto Scaling Notification Access (Acesso à notificação do EC2 Auto Scaling).
   f. Escolha Next (Próximo) duas vezes para ir até a página Name, review, and create (Nomear, revisar e criar).
   g. Em Role name (Nome da função), insira um nome para a função (por exemplo, **my-notification-role**) e escolha Create role (Criar função).
   h. Na página Roles (Funções), escolha a função recém-criada para abrir a página Summary (Resumo). Anote o Role ARN (ARN da função). Por exemplo, `.arn:aws:iam::123456789012:role/my-notification-role` Você precisa dele para criar o gancho do ciclo de vida.

4. Depois que você tiver seguido estas instruções, prossiga para Adicionar ganchos do ciclo de vida (AWS CLI) como próxima etapa.

Encaminhe notificações AWS Lambda diretamente para

É possível usar uma função do Lambda como destino de notificação quando ocorrer uma ação do ciclo de vida.

Para encaminhar notificações AWS Lambda diretamente para

1. Abra a página Functions (Funções) no console do Lambda.
2. Escolha a função do Lambda que você deseja.

   Se você quiser criar uma nova função do Lambda, consulte Criar a função do Lambda

3. Escolha a guia Configuração e, em seguida, Permissões.

4. Role para baixo até Resource-based policy (Política baseada em recurso) e, em seguida, escolha Add permissions (Adicionar permissões). Uma política baseada em recurso é usada para conceder permissões para invocar sua função no principal que é especificado na política. Neste caso, o principal será o perfil vinculado ao serviço do Amazon EC2 Auto Scaling que está associada ao grupo do Auto Scaling.
5. Na seção Policy statement (Declaração da política), configure suas permissões:
   a. Selecione Conta da AWS.
   b. Em Principal insira o ARN da função vinculada a serviço de chamada, por exemplo, **`arn:aws:iam::<aws-account-id>:role/aws-service-role/autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling`**.
   c. Para Ação, escolha lambda: InvokeFunction.
   d. Em Statement ID (ID da instrução), insira um ID de instrução exclusivo, como **`AllowInvokeByAutoScaling`**.
   e. Escolha Salvar.
6. Depois que você tiver seguido estas instruções, prossiga para Adicionar ganchos do ciclo de vida (AWS CLI) como próxima etapa.

Exemplo de mensagem de notificação

Esta seção fornece um exemplo de notificação para Amazon SNS, Amazon SQS e. AWS Lambda

Enquanto a instância está em estado de espera, uma mensagem é publicada no Amazon SNS, no Amazon SQS e no destino da notificação. AWS Lambda

A mensagem inclui as seguintes informações:

- `Origin`: um local de onde está vindo uma instância do EC2.
- `Destination`: um local para onde uma instância do EC2 está indo.
- `LifecycleActionToken` — O token da ação do ciclo de vida.
- `AccountId`— O Conta da AWS ID.
- `AutoScalingGroupName`: o nome do grupo do Auto Scaling.
- `LifecycleHookName` — O nome do gancho do ciclo de vida.
- `EC2InstanceId` — A ID da instância EC2.
- `LifecycleTransition` — O tipo de gancho do ciclo de vida.

- `NotificationMetadata`: os metadados da notificação.

Veja a seguir um exemplo de mensagem de notificação.

```
Service: AWS Auto Scaling
Time: 2021-01-19T00:36:26.533Z
RequestId: 18b2ec17-3e9b-4c15-8024-ff2e8ce8786a
Origin: EC2
Destination: AutoScalingGroup
LifecycleActionToken: 71514b9d-6a40-4b26-8523-05e7ee35fa40
AccountId: 123456789012
AutoScalingGroupName: my-asg
LifecycleHookName: my-hook
EC2InstanceId: i-0598c7d356eba48d7
LifecycleTransition: autoscaling:EC2_INSTANCE_LAUNCHING
NotificationMetadata: hook message metadata
```

Exemplo de mensagem de notificação de teste

Quando você adiciona um gancho do ciclo de vida, uma mensagem de notificação de teste é publicada no destino de notificação. Veja a seguir um exemplo de mensagem de notificação de teste.

```
Service: AWS Auto Scaling
Time: 2021-01-19T00:35:52.359Z
RequestId: 18b2ec17-3e9b-4c15-8024-ff2e8ce8786a
Event: autoscaling:TEST_NOTIFICATION
AccountId: 123456789012
AutoScalingGroupName: my-asg
AutoScalingGroupARN: arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:042cba90-
ad2f-431c-9b4d-6d9055bcc9fb:autoScalingGroupName/my-asg
```

> **Note**
>
> Para obter exemplos dos eventos entregues pelo Amazon EC2 Auto EventBridge Scaling para, consulte. Referência de eventos do Amazon EC2 Auto Scaling

Controle a retenção de instâncias com políticas de ciclo de vida de instâncias

As políticas de ciclo de vida da instância oferecem proteção contra rescisões do Amazon EC2 Auto Scaling quando uma ação do ciclo de vida de rescisão é abandonada. Diferentemente dos ganchos de ciclo de vida isolados, as políticas de ciclo de vida das instâncias são projetadas para garantir que as instâncias sejam transferidas para um estado retido quando procedimentos de desligamento regulares não forem concluídos com êxito.

Quando usar políticas de ciclo de vida da instância

Use políticas de ciclo de vida da instância quando o desligamento normal do seu aplicativo não for opcional, mas obrigatório e desligamentos malsucedidos exigirem intervenção manual. Entre os casos de uso comuns estão:

- Aplicativos com estado sólido que devem concluir a persistência dos dados antes do encerramento.
- Aplicações que exigem períodos de drenagem prolongados que podem exceder o tempo limite máximo do ciclo de vida útil do gancho de 48 horas.
- Cargas de trabalho que lidam com dados confidenciais em que a limpeza falha ou incompleta pode resultar em perda ou corrupção de dados.
- Serviços essenciais em que o desligamento abrupto causa impacto na disponibilidade.

Para obter mais informações sobre como lidar normalmente com o encerramento de instâncias, consulte. Projete suas aplicações para lidar com a terminação de instâncias sem problemas

Como as políticas de ciclo de vida da instância funcionam com ganchos do ciclo de vida de terminação

As políticas de ciclo de vida da instância funcionam em combinação com ganchos do ciclo de vida de terminação, não como substitutas. O processo segue várias etapas:

1. As ações do ciclo de vida de rescisão são executadas. Quando o Amazon EC2 Auto Scaling seleciona uma instância para encerramento, seus ganchos do ciclo de vida de rescisão são invocados e a instância `Terminating:Wait` entra no estado para começar a executar as ações do ciclo de vida de rescisão.

2. Começa uma tentativa graciosa de desligamento. Seu aplicativo, executado na instância ou por meio de um plano de controle, recebe a notificação de ação do ciclo de vida de encerramento e inicia procedimentos de desligamento simples, como drenar conexões, concluir trabalhos em andamento ou transferir dados.
3. As ações do ciclo de vida de rescisão foram concluídas. Uma ação do ciclo de vida de rescisão pode ser concluída com `CONTINUE` ou `ABANDON` resultar.
4. A política de ciclo de vida da instância avalia a situação. Sem uma política de ciclo de vida da instância configurada, a instância prossegue com o encerramento imediatamente, mesmo que a ação do ciclo de vida de encerramento tenha sido concluída com o resultado. `ABANDON` Com uma política de ciclo de vida da instância configurada para reter instâncias ativadas`TerminateHookAbandon`, a instância passa para um estado retido se a ação do ciclo de vida de encerramento for concluída com o resultado. `ABANDON`
5. As instâncias retidas aguardam ação manual. Instâncias em estados retidos continuam incorrendo em cobranças padrão do Amazon EC2. Essas instâncias não contam para a capacidade desejada do seu grupo de Auto Scaling, então o Auto Scaling lança instâncias substitutas para manter o tamanho desejado. Os recursos do Auto Scaling, como atualização da instância e vida útil máxima da instância, também ignorarão as instâncias retidas. Isso permite que você conclua os procedimentos de limpeza manualmente, recupere dados ou investigue por que o desligamento automático falhou antes de encerrar manualmente a instância.
6. A rescisão manual ocorre. Depois de concluir as ações necessárias na instância retida, você precisa chamar a `TerminateInstanceInAutoScalingGroup` API para encerrar a instância.

Configurar a retenção de instâncias

Configure seu grupo Amazon EC2 Auto Scaling para reter instâncias quando as ações do ciclo de vida de encerramento falharem.

Para usar políticas de ciclo de vida da instância em seu grupo de Auto Scaling, você também deve configurar um gancho de ciclo de vida de encerramento. Se você configurar uma política de ciclo de vida da instância, mas não tiver nenhum gancho de ciclo de vida de encerramento, a política não terá efeito. As políticas de ciclo de vida da instância só serão aplicadas quando as ações do ciclo de vida de encerramento forem abandonadas, não quando forem concluídas com êxito com o resultado. `CONTINUE`

As políticas de ciclo de vida da instância usam gatilhos de retenção para determinar quando reter uma instância. O `TerminateHookAbandon` gatilho causa retenção em vários cenários:

- Quando você chama explicitamente a CompleteLifecycleActionAPI com o ABANDON resultado.
- Quando uma ação do ciclo de vida de encerramento com resultado ABANDON padrão expira porque o tempo limite da pulsação é atingido sem receber uma pulsação.
- Quando o tempo limite global é atingido em uma ação do ciclo de vida de encerramento com resultado padrãoABANDON, que é 48 horas ou 100 vezes o tempo limite do batimento cardíaco, o que for menor

Console

Para configurar a retenção de instâncias

1. Abra o console do Amazon EC2 Auto Scaling
2. Crie seu grupo de Auto Scaling (o padrão da política de ciclo de vida da instância é Terminate)
3. Acesse a página de detalhes do seu grupo de Auto Scaling e escolha a guia Gerenciamento de instâncias
4. Em Política de ciclo de vida da instância para ganchos de ciclo de vida, escolha Retain
5. Crie seus ganchos do ciclo de vida da rescisão com:
   - Transição do ciclo de vida definida para encerramento da instância
   - Resultado padrão definido como Abandono

AWS CLI

Para configurar a retenção de instâncias

Use o create-auto-scaling-groupcomando com uma política de ciclo de vida da instância:

```
aws autoscaling create-auto-scaling-group \
--auto-scaling-group-name my-asg \
--launch-template LaunchTemplateName=my-template,Version='$Latest' \
--min-size 1 \
--max-size 3 \
--desired-capacity 2 \
--vpc-zone-identifier subnet-12345678 \
--instance-lifecycle-policy file://lifecycle-policy.json
```

Conteúdo do lifecycle-policy.json:

```
{
    "RetentionTriggers": {
        "TerminateHookAbandon": "retain"
    }
}
```

Para adicionar um gancho do ciclo de vida de término

Use o comando put-lifecycle-hook:

```
aws autoscaling put-lifecycle-hook \
--lifecycle-hook-name my-termination-hook \
--auto-scaling-group-name my-asg \
--lifecycle-transition autoscaling:EC2_INSTANCE_TERMINATING \
--default-result ABANDON \
--heartbeat-timeout 300
```

Gerencie instâncias retidas

Monitore e controle as instâncias do Amazon EC2 que foram movidas para um estado retido. Use CloudWatch métricas para rastrear instâncias retidas e, em seguida, encerre manualmente as instâncias retidas após concluir suas ações personalizadas.

As instâncias retidas não contam para a capacidade desejada do seu grupo Amazon EC2 Auto Scaling. Quando uma instância entra em um estado retido, o Auto Scaling inicia uma instância substituta para manter a capacidade desejada. Por exemplo, suponha que seu grupo de Auto Scaling tenha uma capacidade desejada de 10. Quando uma instância entra no `Terminating:Retained` estado, o Auto Scaling inicia uma instância substituta para manter a capacidade desejada de 10. Agora você tem 11 instâncias em execução no total: 10 em seu grupo ativo mais 1 instância retida. As cobranças padrão do Amazon EC2 para todas as 11 instâncias serão aplicadas até que você encerre manualmente a instância retida.

Estados do ciclo de vida das instâncias retidas

Entenda como as instâncias passam pelos estados do ciclo de vida quando as políticas de ciclo de vida da instância são usadas. As instâncias seguem um caminho específico desde a rescisão normal, passando pela retenção, até a rescisão final.

Quando a retenção é acionada, as instâncias passam pelos seguintes estados:

1. `Terminating`- A rescisão normal começa
2. `Terminating:Wait`- O gancho do ciclo de vida é executado
3. `Terminating:Proceed`- Conclusão das ações do ciclo de vida (sejam elas bem-sucedidas ou fracassadas)
4. `Terminating:Retained`- Falha no gancho, instância retida para intervenção manual

As instâncias de pool quente seguem caminhos de estado de ciclo de vida diferentes, dependendo do cenário:

Instâncias voltando para a piscina aquecida:

1. `Warmed:Pending`- Começa a transição normal da piscina quente
2. `Warmed:Pending:Wait`- O gancho do ciclo de vida é executado
3. `Warmed:Pending:Proceed`- Conclusão das ações do ciclo de vida (sejam elas bem-sucedidas ou fracassadas)
4. `Warmed:Pending:Retained`- Falha no gancho, instância retida para intervenção manual

Instâncias sendo encerradas da piscina aquecida:

1. `Warmed:Terminating`- A rescisão normal começa
2. `Warmed:Terminating:Wait`- O gancho do ciclo de vida é executado
3. `Warmed:Terminating:Proceed`- Conclusão das ações do ciclo de vida (sejam elas bem-sucedidas ou fracassadas)
4. `Warmed:Terminating:Retained`- Falha no gancho, instância retida para intervenção manual

Monitore instâncias retidas

Como as instâncias retidas do Amazon EC2 incorrem em custos e exigem intervenção manual, monitorá-las é essencial. O Amazon EC2 Auto Scaling CloudWatch fornece várias métricas para rastrear instâncias retidas.

Ative as métricas de grupo para rastrear instâncias retidas:

```
aws autoscaling enable-metrics-collection \
--auto-scaling-group-name my-asg \
--metrics GroupTerminatingRetainedInstances
```

As métricas disponíveis são:

- `GroupTerminatingRetainedInstances`mostra o número de instâncias no `Terminating:Retained` estado.
- `GroupTerminatingRetainedCapacity`mostra as unidades de capacidade representadas pelas instâncias no `Terminating:Retained` estado.
- `WarmPoolTerminatingRetainedCapacity`rastreia instâncias retidas que terminam na piscina aquecida.
- `WarmPoolPendingRetainedCapacity`rastreia instâncias retidas retornando à piscina aquecida.

Você também pode verificar as atividades de escalabilidade do seu grupo Amazon EC2 Auto Scaling para entender por que as instâncias foram retidas. Procure atividades de encerramento `StatusCode: Cancelled` e mensagens de motivo de status indicando falhas no gancho do ciclo de vida:

```
aws autoscaling describe-scaling-activities \
--auto-scaling-group-name my-asg
```

Recomendamos criar CloudWatch alarmes sobre essas métricas para alertá-lo quando as instâncias entrarem em um estado retido. Isso ajuda você a monitorar as implicações de custo e garante que você não se esqueça de limpar as instâncias que exigem intervenção manual.

Encerrar instâncias retidas

Depois de concluir suas ações personalizadas, encerre suas instâncias retidas chamando a TerminateInstanceInAutoScalingGroupAPI:

```
aws autoscaling terminate-instance-in-auto-scaling-group \
--instance-id i-1234567890abcdef0 \
--no-should-decrement-desired-capacity
```

Recuperar o estado de destino do ciclo de vida por meio de metadados de instância

Cada instância do Auto Scaling que você inicia passa por vários estados do ciclo de vida. Para invocar ações personalizadas de dentro de uma instância que atuem em transições específicas de estado do ciclo de vida, é necessário recuperar o estado do ciclo de vida de destino por meio de metadados da instância.

Por exemplo, você pode precisar de um mecanismo para detectar o encerramento da instância de dentro da instância para executar algum código na instância antes que ela seja encerrada. Você pode fazer isso escrevendo um código que pesquise o estado do ciclo de vida de uma instância diretamente da instância. Em seguida, será possível adicionar um gancho do ciclo de vida ao grupo do Auto Scaling para manter a instância em execução até que seu código envie o comando complete-lifecycle-action para continuar.

O ciclo de vida de instância do Auto Scaling tem dois estados estáveis primários (`InService` e `Terminated`) e dois estados estáveis paralelos (`Detached` e `Standby`). Se você usar o pool ativo, o ciclo de vida tem mais quatro estados estáveis (`Warmed:Hibernated`, `Warmed:Running`, `Warmed:Stopped` e `Warmed:Terminated`).

Quando uma instância se prepara para fazer a transição para um dos estados estáveis anteriores, o Amazon EC2 Auto Scaling atualiza o valor do item de metadados `autoscaling/target-lifecycle-state` da instância. Para obter o estado do ciclo de vida de destino na instância, é necessário usar o Serviço de Metadados da Instância para recuperá-lo dos metadados da instância.

> **Note**
>
> Os metadados da instância são dados sobre uma instância do Amazon EC2 que as aplicações podem usar para consultar informações de instância. O Instance Metadata Service é um componente na instância que o código local usa para acessar os metadados da instância. O código local pode incluir scripts de dados de usuário ou aplicações em execução na instância.

O código local pode acessar os metadados da instância de uma instância em execução usando um dos dois métodos: Instance Metadata Service Version 1 (IMDSv1) ou Instance Metadata Service Version 2 (). IMDSv2 IMDSv2usa solicitações orientadas à sessão e atenua vários tipos de vulnerabilidades que poderiam ser usadas para tentar acessar os metadados da instância. Para obter detalhes sobre esses dois métodos, consulte Uso IMDSv2 no Guia do usuário do Amazon EC2.

IMDSv2

```
[ec2-user ~]$ TOKEN=`curl -X PUT "http://169.254.169.254/latest/api/token" -H "X-aws-ec2-metadata-token-ttl-seconds: 21600"` \
&& curl -H "X-aws-ec2-metadata-token: $TOKEN" -v http://169.254.169.254/latest/meta-data/autoscaling/target-lifecycle-state
```

IMDSv1

```
[ec2-user ~]$ curl http://169.254.169.254/latest/meta-data/autoscaling/target-lifecycle-state
```

O seguinte é um exemplo de saída.

```
InService
```

O estado de destino do ciclo de vida é o estado para o qual a instância está fazendo a transição. O estado atual do ciclo de vida é o estado no qual a instância está na ocasião. Eles podem ser iguais após a conclusão da ação do ciclo de vida e depois que a instância terminar sua transição para o estado de destino do ciclo de vida. Não é possível recuperar o estado atual do ciclo de vida da instância nos metadados da instância.

Em 10 de março de 2022, o Amazon EC2 Auto Scaling começou a gerar o estado de destino do ciclo de vida. Se sua instância fizer a transição para um dos estados de destino do ciclo de vida após essa data, o item de estado de destino do ciclo de vida estará presente nos metadados de sua instância. Caso contrário, ele não estará presente e você receberá um erro HTTP 404.

Para obter mais informações sobre a recuperação de metadados de instância, consulte Recuperar metadados de instância no Guia do usuário do Amazon EC2.

Para um tutorial que mostra como criar um gancho do ciclo de vida com uma ação personalizada em um script de dados de usuário que usa o estado de destino do ciclo de vida, consulte Tutorial: Usar script de dados e metadados de instância para recuperar o estado do ciclo de vida.

> **Important**
>
> Para garantir que você possa invocar uma ação personalizada o mais rápido possível, seu código local deve pesquisar o IMDS com frequência e repetir os erros.

Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling

Para colocar suas instâncias do Auto Scaling em um estado de espera e executar ações personalizadas nelas, você pode adicionar ganchos do ciclo de vida ao seu grupo do Auto Scaling.

Ações personalizadas são executadas à medida que as instâncias são iniciadas ou antes de serem terminadas. As instâncias permanecem em um estado de espera até que você conclua a ação do ciclo de vida, ou até o período de tempo limite terminar.

Depois de criar um grupo de Auto Scaling a partir do Console de gerenciamento da AWS, você pode adicionar um ou mais ganchos de ciclo de vida a ele, até um total de 50 ganchos de ciclo de vida. Você também pode usar o AWS CLI, CloudFormation, ou um SDK para adicionar ganchos de ciclo de vida a um grupo de Auto Scaling ao criá-lo.

Por padrão, quando você adiciona um gancho de ciclo de vida no console, o Amazon EC2 Auto Scaling envia notificações de eventos de ciclo de vida para a Amazon. EventBridge Usar EventBridge ou usar um script de dados do usuário é uma prática recomendada. Para criar um gancho de ciclo de vida que envie notificações diretamente para o Amazon SNS, o Amazon SQS, AWS Lambda ou você pode usar put-lifecycle-hooko comando, conforme mostrado nos exemplos deste tópico.

Conteúdo

- Adicionar ganchos do ciclo de vida (console)
- Adicionar ganchos do ciclo de vida (AWS CLI)

Adicionar ganchos do ciclo de vida (console)

Siga estas etapas para adicionar ganchos do ciclo de vida ao seu grupo do Auto Scaling. Para adicionar ganchos do ciclo de vida para expansão (inicialização de instâncias) e expansão (terminação de instâncias ou retorno para um pool ativo), é necessário criar dois ganchos separados.

Antes de começar, confirme se você configurou uma ação personalizada, conforme necessário, conforme descrito em Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling.

Para adicionar um gancho do ciclo de vida para expansão

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling. Um painel dividido é aberto na parte inferior da página.
3. Na guia Instance management (Gerenciamento de instâncias), em Lifecycle ganchos (Ganchos do ciclo de vida), escolha Create lifecycle gancho (Criar gancho do ciclo de vida).

4. Para definir um gancho do ciclo de vida para expansão (execução de instâncias), faça o seguinte:

   a. Em Lifecycle Hook Name (Nome do gancho do ciclo de vida), especifique um nome para o gancho do ciclo de vida.

   b. Em Lifecycle transition (Transição do ciclo de vida), escolha Instance launch (Início da instância).

   c. Para Tempo limite de pulsação, especifique o tempo, em segundos, para que as instâncias permaneçam em estado de espera durante a expansão antes que o hook expire. O intervalo é de `30` a `7200` segundos. Definir um longo período de tempo limite fornece mais tempo para a conclusão da sua ação personalizada. Em seguida, se você terminar antes que o período de tempo limite termine, envie o complete-lifecycle-actioncomando para permitir que a instância prossiga para o próximo estado.

   d. Em Resultado padrão, especifique a ação a ser adotada mediante o término do tempo limite do gancho do ciclo de vida ou quando houver uma falha inesperada. Você pode escolher CONTINUAR ou ABANDONAR.

      - Se você escolher CONTINUAR, o grupo do Auto Scaling poderá prosseguir com qualquer outro gancho do ciclo de vida e depois colocar a instância em serviço.
      - Se você escolher ABANDON, o grupo do Auto Scaling interromperá todas as ações restantes e encerrará a instância imediatamente.

   e. (Opcional) Em Metadados de notificação, especifique outras informações que você deseja incluir quando o Amazon EC2 Auto Scaling enviar uma mensagem ao destino da notificação.

5. Escolha Criar.

Para adicionar um gancho do ciclo de vida para redução

1. Escolha Criar gancho do ciclo de vida para continuar de onde você parou depois de criar um gancho do ciclo de vida para expansão.
2. Para definir um gancho do ciclo de vida para redução (instâncias que terminam ou retornam a um pool ativo), faça o seguinte:

   a. Em Lifecycle Hook Name (Nome do gancho do ciclo de vida), especifique um nome para o gancho do ciclo de vida.

b. Em Lifecycle transition (Transição do ciclo de vida), escolha Instance terminate (Término da instância).

c. Para Tempo limite de pulsação, especifique o tempo, em segundos, para que as instâncias permaneçam em estado de espera durante a expansão antes que o hook expire. Recomendamos um curto período de tempo limite de `30` até `120` segundos, dependendo de quanto tempo você precisa para realizar qualquer tarefa final, como extrair registros do EC2. CloudWatch

d. Em Default result (Resultado padrão), especifique a ação que o grupo do Auto Scaling executa quando o tempo limite se esgota ou quando há uma falha inesperada. ABANDON (ABANDONAR) e CONTINUE (CONTINUAR) permitem que a instância termine.

   - Se você escolher CONTINUE (CONTINUAR), o grupo do Auto Scaling poderá prosseguir com todas as ações restantes, como outros ganchos do ciclo de vida, antes do término.
   - Se você escolher ABANDON, o grupo do Auto Scaling encerrará a instância imediatamente.

e. (Opcional) Em Metadados de notificação, especifique outras informações que você deseja incluir quando o Amazon EC2 Auto Scaling enviar uma mensagem ao destino da notificação.

3. Escolha Criar.

Adicionar ganchos do ciclo de vida (AWS CLI)

Crie e atualize ganchos de ciclo de vida usando o comando put-lifecycle-hook.

Para executar uma ação de expansão, use o seguinte comando:

```
aws autoscaling put-lifecycle-hook --lifecycle-hook-name my-launch-hook  \
  --auto-scaling-group-name my-asg \
  --lifecycle-transition autoscaling:EC2_INSTANCE_LAUNCHING
```

Para executar uma ação de redução, use o comando a seguir:

```
aws autoscaling put-lifecycle-hook --lifecycle-hook-name my-termination-hook  \
  --auto-scaling-group-name my-asg \
  --lifecycle-transition autoscaling:EC2_INSTANCE_TERMINATING
```

Para receber notificações usando o Amazon SNS ou o Amazon SQS, adicione as opções `--notification-target-arn` e `--role-arn`. Para receber notificações usando AWS Lambda, adicione `--notification-target-arn` o.

O exemplo a seguir cria um gancho do ciclo de vida que especifica um tópico do SNS chamado*my-sns-topic* como destino de notificação.

```
aws autoscaling put-lifecycle-hook --lifecycle-hook-name my-termination-hook  \
  --auto-scaling-group-name my-asg \
  --lifecycle-transition autoscaling:EC2_INSTANCE_TERMINATING \
  --notification-target-arn arn:aws:sns:region:123456789012:my-sns-topic \
  --role-arn arn:aws:iam::123456789012:role/my-notification-role
```

O tópico recebe uma notificação de teste com o seguinte par de chave/valor:

```
"Event": "autoscaling:TEST_NOTIFICATION"
```

Por padrão, o put-lifecycle-hookcomando cria um gancho de ciclo de vida com um tempo limite de pulsação de `3600` segundos (uma hora).

Para alterar o tempo limite de pulsação de um hook existente do ciclo de vida, adicione a opção `--heartbeat-timeout`, conforme exibido no exemplo a seguir.

```
aws autoscaling put-lifecycle-hook --lifecycle-hook-name my-termination-hook \
  --auto-scaling-group-name my-asg --heartbeat-timeout 120
```

Se uma instância já estiver em estado de espera, você pode evitar que o gancho do ciclo de vida atinja o tempo limite gravando uma pulsação usando o comando CLI record-lifecycle-action-heartbeat. Isso estende o tempo limite pelo valor especificado quando você criou o gancho do ciclo de vida. Se você terminar antes do término do período de tempo limite, poderá enviar o comando da complete-lifecycle-actionCLI para permitir que a instância prossiga para o próximo estado. Para ter mais informações e exemplos, consulte Concluir uma ação do ciclo de vida em um grupo do Auto Scaling.

Concluir uma ação do ciclo de vida em um grupo do Auto Scaling

Quando o grupo do Auto Scaling responde a um evento do ciclo de vida, ele coloca a instância em um estado de espera e envia notificação de evento. Enquanto a instância está em estado de espera, você pode executar uma ação personalizada.

Concluir a ação do ciclo de vida com um resultado de `CONTINUE` é útil se você terminar antes que o tempo limite expire. Se você não concluir a ação do ciclo de vida, o gancho do ciclo de vida vai para o status que você especificou para Resultado padrão após o término do tempo limite.

Conteúdo

- Concluir uma ação do ciclo de vida (manual)
- Concluir uma ação do ciclo de vida (automática)

Concluir uma ação do ciclo de vida (manual)

O procedimento a seguir é para a interface de linha de comando e não tem suporte ao console. Informações que devem ser substituídas, como o ID da instância ou o nome de um grupo do Auto Scaling, são mostradas em itálico.

Para concluir uma ação do ciclo de vida (AWS CLI)

1. Se você precisar de mais tempo para concluir a ação personalizada, use o comando record-lifecycle-action-heartbeat para reiniciar o período de tempo limite e manter a instância em estado de espera. Por exemplo, se o período de tempo limite for 1 hora e você chamar esse comando após 30 minutos, a instância permanecerá em estado de espera por mais 1 hora ou por um total de 90 minutos.

   Você pode especificar o token de ação do ciclo de vida recebido com a notificação, conforme é mostrado no comando a seguir.

   ```
   aws autoscaling record-lifecycle-action-heartbeat --lifecycle-hook-name my-launch-hook \
     --auto-scaling-group-name my-asg --lifecycle-action-token bcd2f1b8-9a78-44d3-8a7a-4dd07d7cf635
   ```

   Como alternativa, é possível especificar o ID da instância, recebido com a notificação, conforme mostrado no comando a seguir.

   ```
   aws autoscaling record-lifecycle-action-heartbeat --lifecycle-hook-name my-launch-hook \
     --auto-scaling-group-name my-asg --instance-id i-1a2b3c4d
   ```

2. Se você concluir a ação personalizada antes que o período de tempo limite termine, use o complete-lifecycle-actioncomando para que o grupo do Auto Scaling possa continuar iniciando

ou encerrando a instância. Você pode especificar o token da ação do ciclo de vida, conforme mostrado no comando a seguir:

```
aws autoscaling complete-lifecycle-action --lifecycle-action-result CONTINUE \
  --lifecycle-hook-name my-launch-hook --auto-scaling-group-name my-asg \
  --lifecycle-action-token bcd2f1b8-9a78-44d3-8a7a-4dd07d7cf635
```

Como alternativa, você pode especificar o ID da instância, conforme mostrado no comando a seguir:

```
aws autoscaling complete-lifecycle-action --lifecycle-action-result CONTINUE \
  --instance-id i-1a2b3c4d --lifecycle-hook-name my-launch-hook \
  --auto-scaling-group-name my-asg
```

Concluir uma ação do ciclo de vida (automática)

Se você tiver um script de dados do usuário que configure suas instâncias após elas serem iniciadas, você não precisará concluit manualmente as ações do ciclo de vida. Você pode adicionar o complete-lifecycle-actioncomando ao script. O script pode recuperar o ID da instância dos metadados da instância e sinalizar ao Amazon EC2 Auto Scaling quando os scripts de bootstrap tiverem sido concluídos com êxito.

Se você já não estiver fazendo isso, atualize seu script para recuperar o ID da instância nos metadados da instância. Para mais informações, consulte Recuperação de metadados de instância no Guia do usuário do Amazon EC2 para instâncias de Linux.

Se usar o Lambda, você também poderá configurar um retorno de chamada no código da função para permitir que o ciclo de vida da instância prossiga se a ação personalizada tiver êxito. Para obter mais informações, consulte Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda.

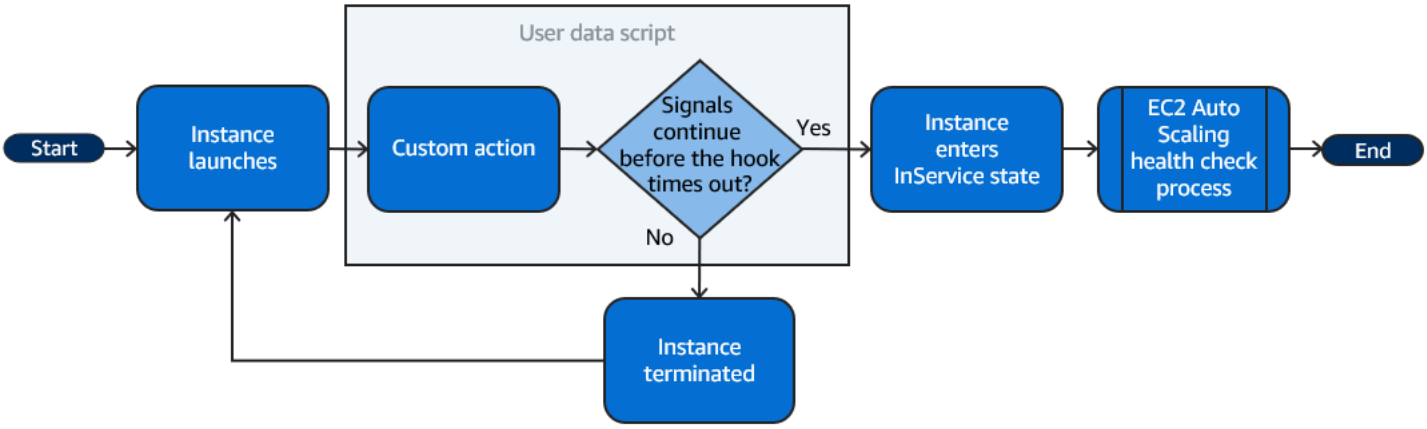
Tutorial: Usar script de dados e metadados de instância para recuperar o estado do ciclo de vida

Uma forma comum de criar ações personalizadas para ganchos de ciclo de vida é usar notificações que o Amazon EC2 Auto Scaling envia para outros serviços, como a Amazon. EventBridge No entanto, usando um script de dados do usuário para mover o código que configura instâncias e

conclui a ação do ciclo de vida para as próprias instâncias, você pode evitar a necessidade de criar infraestrutura adicional.

O tutorial a seguir mostra como começar a usar um script de dados do usuário e metadados de instância. Você cria uma configuração básica de grupo do Auto Scaling com um script de dados do usuário que lê o estado de destino do ciclo de vida das instâncias em seu grupo e executa uma ação de retorno de chamada em uma fase específica do ciclo de vida de uma instância para continuar o processo de execução.

A ilustração a seguir resume o fluxo de um evento de aumento a escala horizontalmente quando você usa um script de dados do usuário para executar uma ação personalizada. Depois da execução de uma instância, o ciclo de vida da instância é pausado até que o gancho do ciclo de vida seja concluído, seja por tempo limite ou pelo Amazon EC2 Auto Scaling recebendo um sinal para continuar.



Conteúdo

- Etapa 1: criar um perfil do IAM com permissões para concluir ações do ciclo de vida
- Etapa 2: criar um modelo de execução e incluir o perfil do IAM e um script de dados de usuário
- Etapa 3: criar um grupo do Auto Scaling
- Etapa 4: Adicionar um gancho do ciclo de vida
- Etapa 5: testar e verificar a funcionalidade
- Etapa 6: limpar
- Recursos relacionados

Etapa 1: criar um perfil do IAM com permissões para concluir ações do ciclo de vida

Ao usar o AWS CLI ou um AWS SDK para enviar um retorno de chamada para concluir as ações do ciclo de vida, você deve usar uma função do IAM com permissões para concluir as ações do ciclo de vida.

Para criar a política

1. Abra a página Policies (Políticas) do console do IAM e escolha Create policy (Criar política).
2. Selecione a guia JSON.
3. Na caixa Policy Document (Documento de política), copie e cole o seguinte documento de política. *sample text*Substitua o pelo número da sua conta e pelo nome do grupo Auto Scaling que você deseja criar ()**TestAutoScalingEvent-group**.

   JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "autoscaling:CompleteLifecycleAction"
      ],
      "Resource":
 "arn:aws:autoscaling:*:123456789012:autoScalingGroup:*:autoScalingGroupName/
TestAutoScalingEvent-group"
    }
  ]
}
```

4. Escolha Próximo.
5. Em Nome da política, insira **TestAutoScalingEvent-policy**. Selecione Criar política.

Quando você terminar de criar a política, poderá criar uma função que a utilize.

Para criar o perfil

1. No painel de navegação à esquerda, escolha Roles (Funções).

2. Selecione Criar perfil.
3. Em Select trusted entity (Selecionar entidade confiável), escolha AWS Service (Serviço).
4. Para seu caso de uso, escolha EC2 e escolha Next (Próximo).
5. Em Adicionar permissões, escolha a política que você criou (TestAutoScalingEvent-policy). Em seguida, escolha Próximo.
6. Na página Name, review, and create (Nomear, revisar e criar), em Role name (Nome da função), insira **`TestAutoScalingEvent-role`** e escolha Create role (Criar função).

Etapa 2: criar um modelo de execução e incluir o perfil do IAM e um script de dados de usuário

Crie um modelo de execução para usar com seu grupo do Auto Scaling. Inclua o perfil do IAM que você criou e a amostra de script de dados do usuário fornecida.

Para criar um modelo de execução

1. Abra a página Launch templates (Modelos de execução) do console do Amazon EC2.
2. Escolha Criar modelo de execução.
3. Para o Launch template name (Nome do modelo de execução), insira **`TestAutoScalingEvent-template`**.
4. Em Auto Scaling guidance (Guia do Auto Scaling), marque a caixa de seleção.
5. Para Imagens de aplicação e SO (Amazon Machine Image), escolha Amazon Linux 2 (HVM), SSD Volume Type, 64 bits (x86) na lista Quick Start (Início rápido).
6. Em Instance type (Tipo de instância), escolha um tipo de instância do Amazon EC2 (p. ex., “t2.micro”).
7. Em Advanced details (Detalhes avançados), expanda a seção para visualizar os campos.
8. Para o perfil da instância do IAM, escolha o nome do perfil da instância do IAM da sua função do IAM (TestAutoScalingEvent-role). Um perfil de instância é um contêiner para um perfil do IAM que permite ao Amazon EC2 passar o perfil do IAM para uma instância quando ela é iniciada.

   Se tiver criado um perfil do IAM usando o console do IAM, o console terá criado automaticamente um perfil da instância e dará a esse perfil o mesmo nome da função correspondente.
9. Em User data (Dados do usuário), copie e cole a seguinte amostra de script de dados de usuário no campo. Substitua o texto de amostra pelo `group_name` nome do grupo de Auto Scaling

que você deseja criar e `region` pelo que Região da AWS você deseja que seu grupo de Auto Scaling use.

```
#!/bin/bash

function token {
    echo "X-aws-ec2-metadata-token: $(curl -X PUT 'http://169.254.169.254/latest/
api/token' -H 'X-aws-ec2-metadata-token-ttl-seconds: 21600')"
}

function get_target_state {
    echo $(curl -H "$(token)" -s http://169.254.169.254/latest/meta-data/
autoscaling/target-lifecycle-state)
}

function get_instance_id {
    echo $(curl -H "$(token)" -s http://169.254.169.254/latest/meta-data/instance-
id)
}

function complete_lifecycle_action {
    instance_id=$(get_instance_id)
    group_name='TestAutoScalingEvent-group'
    region='us-west-2'

    echo $instance_id
    echo $region
    echo $(aws autoscaling complete-lifecycle-action \
      --lifecycle-hook-name TestAutoScalingEvent-hook \
      --auto-scaling-group-name $group_name \
      --lifecycle-action-result CONTINUE \
      --instance-id $instance_id \
      --region $region)
}

function main {
    while true
    do
        target_state=$(get_target_state)
        if [ \"$target_state\" = \"InService\" ]; then
            # Change hostname
            export new_hostname="${group_name}-$instance_id"
            hostname $new_hostname
```

```
            # Send callback
            complete_lifecycle_action
            break
        fi
        echo $target_state
        sleep 5
    done
}

main
```

Esse script de dados de usuário simples faz o seguinte:

- Chama os metadados da instância para recuperar o estado de destino do ciclo de vida e o ID da instância nos metadados da instância
- Recupera o estado de destino do ciclo de vida repetidamente até que ele mude para `InService`
- Altera o nome de host da instância para o ID da instância tendo como prefixo o nome do grupo do Auto Scaling, se o estado de destino do ciclo de vida for `InService`
- Envia um retorno de chamada chamando o comando complete-lifecycle-action da CLI para sinalizar o Amazon EC2 Auto Scaling a `CONTINUE` o processo de execução do EC2

10. Escolha Criar modelo de execução.
11. Na página de confirmação, escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

> **Note**
>
> Para outros exemplos que você pode usar como referência para desenvolver seu script de dados de usuário, consulte o GitHub repositório do Amazon EC2 Auto Scaling.

Etapa 3: criar um grupo do Auto Scaling

Depois de criar seu modelo de execução, crie um grupo do Auto Scaling.

Para criar um grupo do Auto Scaling

1. Na página Choose launch template or configuration (Escolher o modelo ou a configuração de execução), em Auto Scaling group name (Nome do grupo do Auto Scaling), insira um nome para o grupo do Auto Scaling (**`TestAutoScalingEvent-group`**).
2. Escolha Next (Próximo) e acesse a página Choose instance launch options (Escolher as opções de execução de instância).
3. Em Network (Rede), selecione uma VPC.
4. Em Availability Zones and subnets (Zonas de disponibilidade e sub-redes), escolha uma ou mais sub-redes de uma ou mais zonas de disponibilidade.
5. Na seção Instance type requirements (Requisitos de tipo de instância), use a configuração padrão para simplificar essa etapa. (Não substitua o modelo de execução.) Neste tutorial, você fará o execução de apenas uma das Instâncias sob demanda usando o tipo de instância especificado no modelo de execução.
6. Selecione Skip to review (Pular para a revisão) na parte inferior da tela.
7. Na página Review (Revisar), reveja as configurações do grupo do Auto Scaling e escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Etapa 4: Adicionar um gancho do ciclo de vida

Adicione um gancho do ciclo de vida para manter a instância em um estado de espera até que a ação do ciclo de vida esteja concluída.

Para adicionar um gancho do ciclo de vida

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling. Um painel dividido é aberto na parte inferior da página.
3. No painel inferior, na guia Instance management (Gerenciamento de instâncias), em Lifecycle ganchos (Ganchos do ciclo de vida), escolha Create lifecycle gancho (Criar gancho do ciclo de vida).
4. Para definir um gancho do ciclo de vida para expansão (execução de instâncias), faça o seguinte:

   a. Em Lifecycle gancho name (Nome do gancho do ciclo de vida), insira **`TestAutoScalingEvent-hook`**.

b. Em Lifecycle transition (Transição do ciclo de vida), escolha Instance launch (Início da instância).

c. Em Heartbeat timeout (Tempo limite de pulsação), insira **300** para o número de segundos de espera por um retorno de chamada do seu script de dados de usuário.

d. Em Default result (Resultado padrão), escolha ABANDON (Abandono). Se o hook expirar sem receber um retorno de chamada do script de dados de usuário, o grupo do Auto Scaling encerrará a nova instância.

e. (Opcional) Mantenha Notification metadata (Metadados de notificação) em branco.

5. Escolha Criar.

Etapa 5: testar e verificar a funcionalidade

Para testar a funcionalidade, atualize o grupo do Auto Scaling aumentando em 1 a capacidade desejada do grupo do Auto Scaling. O script de dados de usuário é executado e começa a verificar o estado de destino do ciclo de vida da instância logo após a execução da instância. O script altera o nome do host e envia uma ação de retorno de chamada quando o estado de destino do ciclo de vida for `InService`. Isso geralmente leva apenas alguns segundos para terminar.

Para aumentar o tamanho de grupo do Auto Scalin

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling. Veja os detalhes em um painel inferior enquanto ainda vê as linhas superiores do painel superior.
3. No painel inferior, na guia Details (Detalhes), escolha Group details (Detalhes do grupo, Edit (Editar).
4. Em Desired capacity (Capacidade desejada), aumente o valor atual em 1.
5. Selecione Atualizar. Enquanto a instância está sendo iniciada ou terminada, a coluna Status no painel superior exibe um status Updating capacity (Atualizando capacidade).

Após aumentar a capacidade desejada, você pode verificar na descrição das ações de escalabilidade se sua instância foi executada com êxito e não foi encerrada.

Para visualizar as atividades de escalabilidade

1. Retorne á página Auto Scaling groups (Grupos do Auto Scaling) e selecione seu grupo.

2. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status exibe se o seu grupo do Auto Scaling iniciou uma instância com êxito.
3. Se o script de dados de usuário falhar, você observará uma ação de escalabilidade com um status de `Canceled` e uma mensagem de status de `Instance failed to complete user's Lifecycle Action: Lifecycle Action with token e85eb647-4fe0-4909-b341-a6c42EXAMPLE was abandoned: Lifecycle Action Completed with ABANDON Result` após o término do período de tempo limite.

Etapa 6: limpar

Se tiver terminado de trabalhar com os recursos que criou exclusivamente para este tutorial, siga as etapas abaixo para excluí-los.

Para excluir o gancho do ciclo de vida

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.
3. Na guia Instance management (Gerenciamento de instâncias), em Lifecycle ganchos (Ganchos do ciclo de vida), escolha o gancho do ciclo de vida (`TestAutoScalingEvent-hook`).
4. Selecione Ações, Excluir.
5. Para confirmar, escolha Delete (Excluir) novamente.

Para excluir o modelo de execução

1. Abra a página Launch templates (Modelos de execução) do console do Amazon EC2.
2. Selecione seu modelo de execução (`TestAutoScalingEvent-template`) e escolha Actions (Ações), Delete template (Excluir modelo).
3. Quando a confirmação for solicitada, digite **`Delete`** para confirmar a exclusão do modelo de execução especificado e, em seguida, escolha Excluir.

Se tiver terminado de trabalhar com o grupo de exemplo do Auto Scaling, exclua-o. Você também pode excluir o perfil do IAM e a política de permissões que criou.

Para excluir o grupo do Auto Scaling

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.

2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling (`TestAutoScalingEvent-group`) e escolha Delete (Excluir).
3. Quando a confirmação for solicitada, digite **`delete`** para confirmar a exclusão do grupo do Auto Scaling especificado e, em seguida, escolha Excluir.

   Um ícone de carregamento na coluna Name (Nome) indica que o grupo do Auto Scaling está sendo excluído. É necessário aguardar alguns minutos para encerrar a instância e excluir o grupo.

Para excluir o perfil do IAM

1. Abra a página Roles (Funções) no console do IAM.
2. Selecione o papel da função (`TestAutoScalingEvent-role`).
3. Escolha Excluir.
4. Quando for solicitada confirmação, digite o nome da função e escolha Excluir.

Para excluir a política do IAM

1. Abra a página Policies (Políticas) do console do IAM.
2. Selecione a política que você criou (`TestAutoScalingEvent-policy`).
3. Selecione Ações, Excluir.
4. Quando for solicitada confirmação, digite o nome da política e escolha Excluir.

Recursos relacionados

Os tópicos relacionados a seguir podem ser úteis à medida que você desenvolve um código que invoca ações em instâncias com base nos dados disponíveis nos metadados da instância.

- Recuperar o estado de destino do ciclo de vida por meio de metadados de instância. Esta seção descreve o estado do ciclo de vida de outros casos de uso, como o encerramento da instância.
- Adicionar ganchos do ciclo de vida (console). Este procedimento mostra como adicionar ganchos do ciclo de vida tanto para expansão (execução de instâncias) quanto para redução (instâncias encerrando ou retornando a um pool ativo).

- Categorias de metadados de instância no Guia do usuário do Amazon EC2 para instâncias do Linux. Este tópico lista todas as categorias de metadados de instância que você pode usar para invocar ações em instâncias do EC2.

Para ver um tutorial que mostra como usar a Amazon EventBridge para criar regras que invocam funções Lambda com base em eventos que acontecem com as instâncias em seu grupo de Auto Scaling, consulte. Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda

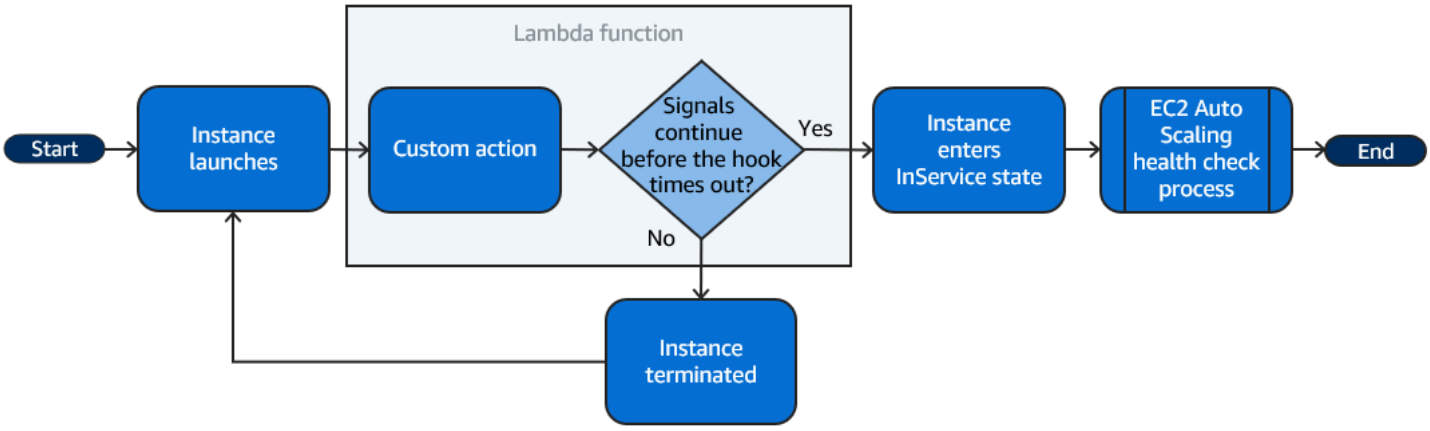
Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda

Neste exercício, você cria uma EventBridge regra da Amazon que inclui um padrão de filtro que, quando combinado, invoca uma AWS Lambda função como destino da regra. Nós fornecemos o padrão de filtro e código de função de exemplo a ser usada.

Se tudo estiver configurado corretamente, no final deste tutorial, a função do Lambda executará uma ação personalizada quando as instâncias forem iniciadas. A ação personalizada simplesmente registra o evento no stream de CloudWatch registros de registros associado à função Lambda.

A função do Lambda também executa um retorno de chamada para permitir que o ciclo de vida da instância prossiga se essa ação for bem-sucedida, mas permite que a instância abandone o início e termine se a ação falhar.

A ilustração a seguir resume o fluxo de um evento de aumento a escala horizontalmente quando você usa uma função do Lambda para executar uma ação personalizada. Depois da execução de uma instância, o ciclo de vida da instância é pausado até que o gancho do ciclo de vida seja concluído, seja por tempo limite ou pelo Amazon EC2 Auto Scaling recebendo um sinal para continuar.



> **Note**
>
> Dependendo do seu caso de uso, você pode configurar um gancho de ciclo de vida seguindo as etapas abaixo e criando uma EventBridge regra. Ou você pode usar uma função Lambda para configurar um gancho de ciclo de vida diretamente sem criar uma regra. EventBridge

Conteúdos

- Pré-requisitos
- Etapa 1: criar um perfil do IAM com permissões para concluir ações do ciclo de vida
- Etapa 2: Criar uma função do Lambda
- Etapa 3: criar uma EventBridge regra
- Etapa 4: Adicionar um gancho do ciclo de vida
- Etapa 5: Testar e verificar o evento
- Etapa 6: limpar
- Recursos relacionados

Pré-requisitos

Antes de iniciar este tutorial, crie um grupo do Auto Scaling, se você ainda não tiver um. Para criar um grupo do Auto Scaling, abra a página Grupos do Auto Scaling do console do Amazon EC2 e escolha Criar grupo de Auto Scaling.

Etapa 1: criar um perfil do IAM com permissões para concluir ações do ciclo de vida

Antes de criar uma função do Lambda, é necessário primeiro criar uma função de execução e uma política de permissões para permitir que o Lambda conclua os ganchos do ciclo de vida.

Para criar a política

1. Abra a página Policies (Políticas) do console do IAM e escolha Create policy (Criar política).
2. Selecione a guia JSON.
3. Na caixa Documento de política, cole o seguinte documento de política na caixa, substituindo o texto pelo número da sua conta e pelo nome do seu grupo de Auto Scaling. *italics*

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "autoscaling:CompleteLifecycleAction"
      ],
      "Resource":
 "arn:aws:autoscaling:*:123456789012:autoScalingGroup:*:autoScalingGroupName/
my-asg"
    }
  ]
}
```

4. Escolha Próximo.
5. Em Nome da política, insira **`LogAutoScalingEvent-policy`**. Selecione Criar política.

Quando você terminar de criar a política, poderá criar uma função que a utilize.

Para criar o perfil

1. No painel de navegação à esquerda, escolha Roles (Funções).
2. Selecione Criar perfil.
3. Em Select trusted entity (Selecionar entidade confiável), escolha AWS service (serviço).
4. Para seu caso de uso, escolha Lambda e escolha Next (Próximo).
5. Em Adicionar permissões, escolha a política que você criou (LogAutoScalingEvent-policy) e a política nomeada AWSLambdaBasicExecutionRole. Em seguida, escolha Próximo.

> **Note**
>
> A AWSLambdaBasicExecutionRolepolítica tem as permissões que a função precisa para gravar registros em CloudWatch Logs.

6. Na página Name, review, and create (Nomear, revisar e criar), em Role name (Nome da função), insira **`LogAutoScalingEvent-role`** e escolha Create role (Criar função).

Etapa 2: Criar uma função do Lambda

Crie uma função do Lambda para servir como destino para eventos. A função Lambda de amostra, escrita em Node.js, é invocada EventBridge quando um evento correspondente é emitido pelo Amazon EC2 Auto Scaling.

Criar uma função do Lambda

1. Abra a página Functions (Funções) no console do Lambda.
2. Escolha Create function (Criar função) e Author from scratch (Criar desde o início).
3. Em Basic information (Informações básicas), em Function name (Nome da função), insira **`LogAutoScalingEvent`**.
4. Em Runtime, selecione Node.js 18.x.
5. Role para baixo e escolha Alterar função de execução padrão e, em seguida, para Função de execução, escolha Usar uma função existente.
6. Em Função existente, escolha LogAutoScalingEvent-role.
7. Deixe os outros valores padrão.
8. Escolha a opção Criar função. Você é retornado ao código e configuração da função.
9. Com sua `LogAutoScalingEvent` função ainda aberta no console, em Código-fonte, no editor, cole o código de exemplo a seguir no arquivo denominado index.mjs.

```
import { AutoScalingClient, CompleteLifecycleActionCommand } from "@aws-sdk/client-
auto-scaling";
export const handler = async(event) => {
  console.log('LogAutoScalingEvent');
  console.log('Received event:', JSON.stringify(event, null, 2));
  var autoscaling = new AutoScalingClient({ region: event.region });
  var eventDetail = event.detail;
  var params = {
    AutoScalingGroupName: eventDetail['AutoScalingGroupName'], /* required */
    LifecycleActionResult: 'CONTINUE', /* required */
    LifecycleHookName: eventDetail['LifecycleHookName'], /* required */
    InstanceId: eventDetail['EC2InstanceId'],
    LifecycleActionToken: eventDetail['LifecycleActionToken']
  };
  var response;
  const command = new CompleteLifecycleActionCommand(params);
  try {
    var data = await autoscaling.send(command);
```

```
    console.log(data); // successful response
    response = {
      statusCode: 200,
      body: JSON.stringify('SUCCESS'),
    };
  } catch (err) {
    console.log(err, err.stack); // an error occurred
    response = {
      statusCode: 500,
      body: JSON.stringify('ERROR'),
    };
  }
  return response;
};
```

Esse código simplesmente registra o evento para que, no final deste tutorial, você possa ver um evento aparecer no stream de CloudWatch registros de registros associado a essa função Lambda.

10. Escolha Implantar.

Etapa 3: criar uma EventBridge regra

Crie uma EventBridge regra para executar sua função Lambda. Para obter mais informações sobre o uso de EventBridge, consulte Use EventBridge para lidar com eventos do Auto Scaling.

Como criar uma regra usando o console

1. Abra o console do EventBridge.
2. No painel de navegação, escolha Regras.
3. Escolha Create rule.
4. Em Define rule detail (Definir detalhe da regra), faça o seguinte:
   a. Em Nome, digite **`LogAutoScalingEvent-rule`**.
   b. Em Barramento de eventos, escolha padrão. Quando um AWS service (Serviço da AWS) em sua conta gera um evento, ele sempre vai para o ônibus de eventos padrão da sua conta.
   c. Em Rule type, escolha Rule with an event pattern.
   d. Escolha Próximo.

5. Em Build event pattern (Criar padrão de evento), faça o seguinte:

   a. Em Origem do evento, escolha AWS eventos ou eventos de EventBridge parceiros.

   b. Role para baixo até o Padrão de eventos e faça o seguinte:

   c. i. Para Origem do evento, escolha Serviços da AWS.

      ii. Em AWS service (Serviço da AWS), escolha Auto Scaling.

      iii. Em Event type (Tipo de evento), selecione Instance Launch and Terminate (Inicialização e encerramento de instância).

      iv. Por padrão, a regra faz a correspondência com qualquer evento de aumento ou redução horizontal da escala. Para criar uma regra que notifique você quando houver um evento de aumento horizontal da escala e uma instância for colocada em estado de espera devido a um gancho do ciclo de vida, escolhaSpecific instance event(s) (Eventos específicos de instância) e selecione EC2 Instance-launch Lifecycle Action (Ação do ciclo de vida de inicialização de instância do EC2).

      v. Por padrão, a regra corresponde a qualquer grupo do Auto Scaling na região. Para fazer com que a regra corresponda a um grupo específico do Auto Scaling, escolha Nome(s) de grupo específico(s) e selecione o grupo.

      vi. Escolha Próximo.

6. Em Select target(s) (Selecionar destino(s)), faça o seguinte:

   a. Em Target types (Tipos de destino), escolha AWS service (Serviço da AWS).

   b. Para Selecionar um destino, escolha Função do Lambda.

   c. Em Função, escolha LogAutoScalingEvent.

   d. Escolha Next (Próximo) duas vezes.

7. Na página Revisar e criar, escolha Criar regra.

Etapa 4: Adicionar um gancho do ciclo de vida

Nesta seção, você adicionará um gancho do ciclo de vida para que o Lambda execute sua função em instâncias no início.

Para adicionar um gancho do ciclo de vida

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.

2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling. Um painel dividido é aberto na parte inferior da página.
3. No painel inferior, na guia Instance management (Gerenciamento de instâncias), em Lifecycle ganchos (Ganchos do ciclo de vida), escolha Create lifecycle gancho (Criar gancho do ciclo de vida).
4. Para definir um gancho do ciclo de vida para expansão (execução de instâncias), faça o seguinte:

   a. Em Lifecycle gancho name (Nome do gancho do ciclo de vida), insira **`LogAutoScalingEvent-hook`**.

   b. Em Lifecycle transition (Transição do ciclo de vida), escolha Instance launch (Início da instância).

   c. Em Heartbeat timeout (Tempo limite de pulsação), insira **`300`** para o número de segundos de espera por um retorno de chamada da sua função do Lambda.

   d. Em Default result (Resultado padrão), escolha ABANDON (Abandono). Isso significa que o grupo do Auto Scaling terminará uma nova instância se o hook expirar sem receber um retorno de chamada de sua função do Lambda.

   e. (Opcional) Deixe Notification metadata (Metados da notificação) vazio. Os dados do evento para os quais passamos EventBridge contêm todas as informações necessárias para invocar a função Lambda.
5. Escolha Criar.

Etapa 5: Testar e verificar o evento

Para testar o evento, atualize o grupo do Auto Scaling aumentando a capacidade desejada do grupo do Auto Scaling em 1. Sua função do Lambda é invocada dentro de alguns segundos depois do aumento da capacidade desejada.

Para aumentar o tamanho de grupo do Auto Scalin

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling para visualizar detalhes em um painel inferior e ainda ver as linhas superiores do painel superior.
3. No painel inferior, na guia Details (Detalhes), escolha Group details (Detalhes do grupo, Edit (Editar).

4. Em Desired capacity (Capacidade desejada), aumente o valor atual em 1.
5. Selecione Atualizar. Enquanto a instância está sendo iniciada ou terminada, a coluna Status no painel superior exibe um status Updating capacity (Atualizando capacidade).

Depois de aumentar a capacidade desejada, você poderá verificar se a sua função do Lambda foi invocada.

Para visualizar a saída da função do Lambda

1. Abra a página Grupos de registros do CloudWatch console.
2. Selecione o nome do grupo de logs para sua função do Lambda (`/aws/lambda/LogAutoScalingEvent`).
3. Selecione o nome do fluxo de logs para visualizar os dados fornecidos pela função para a ação do ciclo de vida.

Em seguida, é possível verificar se a instância foi iniciada com êxito a partir da descrição das atividades de escalabilidade.

Para visualizar as atividades de escalabilidade

1. Retorne á página Auto Scaling groups (Grupos do Auto Scaling) e selecione seu grupo.
2. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status exibe se o seu grupo do Auto Scaling iniciou uma instância com êxito.

   - Se a ação foi bem-sucedida, a atividade de escalabilidade terá o status “Successful” (Sucesso).
   - Se falhar, depois de esperar alguns minutos, você observará uma atividade de escalabilidade com o status “Cancelled” (Cancelado) e uma mensagem de status "Instance failed to complete user's Lifecycle Action: Lifecycle Action with token e85eb647-4fe0-4909-b341-a6c42EXAMPLE was abandoned: Lifecycle Action Completed with ABANDON Result" (Instância falhou ao concluir a ação do ciclo de vida do usuário: ação do ciclo de vida com token e85eb647-4fe0-4909-b341-a6c42EXAMPLE foi abandonada: ação do ciclo de vida concluída com o resultado ABANDONAR).

Para reduzir o tamanho do grupo do Auto Scaling

Se não for necessária a instância adicional iniciada para este teste, você pode abrir a guia Details (Detalhes) e reduzir Desired capacity (Capacidade desejada) em 1.

Etapa 6: limpar

Se você tiver terminado de trabalhar com os recursos que você criou apenas para este tutorial, use as seguintes etapas para excluí-los.

Para excluir o gancho do ciclo de vida

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.
3. Na guia Instance management (Gerenciamento de instâncias), em Lifecycle ganchos (Ganchos do ciclo de vida), escolha o gancho do ciclo de vida (`LogAutoScalingEvent-hook`).
4. Selecione Ações, Excluir.
5. Para confirmar, escolha Delete (Excluir) novamente.

Para excluir a EventBridge regra da Amazon

1. Abra a página de regras no EventBridge console da Amazon.
2. Em Event bus (Barramento de eventos), escolha o barramento de eventos associado à regra (`Default`).
3. Marque a caixa de seleção ao lado da sua regra (`LogAutoScalingEvent-rule`).
4. Escolha Excluir.
5. Quando for solicitada confirmação, digite o nome da regra e escolha Excluir.

Se você tiver terminado de trabalhar com a função de exemplo, exclua-a. Você também pode excluir o grupo de logs que armazena os logs da função e a função de execução e a política de permissões que você criou.

Para excluir uma função do Lambda

1. Abra a página Functions (Funções) no console do Lambda.
2. Escolha a função (`LogAutoScalingEvent`).
3. Selecione Ações, Excluir.

4. Quando for solicitada confirmação, digite **`delete`** para confirmar a exclusão do modelo de execução especificado e, em seguida, escolha Excluir.

Para excluir o grupo de logs

1. Abra a página Grupos de registros do CloudWatch console.
2. Selecione o grupo de logs da função (`/aws/lambda/LogAutoScalingEvent`).
3. Selecione Actions (Ações), Delete log group(s) (Excluir grupo(s) de log).
4. Na caixa de diálogo Delete log group(s) (Excluir grupo(s) de logs), escolha Delete (Excluir).

Para excluir a função de execução

1. Abra a página Roles (Funções) no console do IAM.
2. Selecione o papel da função (`LogAutoScalingEvent-role`).
3. Escolha Excluir.
4. Quando for solicitada confirmação, digite o nome da função e escolha Excluir.

Para excluir a política do IAM

1. Abra a página Policies (Políticas) do console do IAM.
2. Selecione a política que você criou (`LogAutoScalingEvent-policy`).
3. Selecione Ações, Excluir.
4. Quando for solicitada confirmação, digite o nome da política e escolha Excluir.

Recursos relacionados

Os tópicos relacionados a seguir podem ser úteis à medida que você cria EventBridge regras com base em eventos que acontecem nas instâncias do seu grupo de Auto Scaling.

- Use EventBridge para lidar com eventos do Auto Scaling. Esta seção mostra exemplos de eventos para outros casos de uso, incluindo eventos para redução.
- Adicionar ganchos do ciclo de vida (console). Este procedimento mostra como adicionar ganchos do ciclo de vida tanto para expansão (execução de instâncias) quanto para redução (instâncias encerrando ou retornando a um pool ativo).

Para ver um tutorial que mostra como usar o serviço de metadados de instância (IMDS) para invocar uma ação de dentro da própria instância, consulte. Tutorial: Usar script de dados e metadados de instância para recuperar o estado do ciclo de vida

Diminua a latência para aplicações com tempos de inicialização longos usando pools ativos

Um pool ativo oferece a capacidade de diminuir a latência para suas aplicações que apresentam tempos de inicialização excepcionalmente longos, por exemplo, porque as instâncias precisam gravar grandes quantidades de dados no disco. Com os pools ativos, você não precisa mais provisionar excessivamente seus grupos do Auto Scaling para gerenciar a latência a fim de melhorar a performance das aplicações. Para obter mais informações, consulte a postagem do blog Escalabilidade mais rápida de aplicações com pools ativos do EC2 Auto Scaling.

> **Important**
>
> Criar um pool ativo quando ele não é necessário pode gerar custos desnecessários. Se o tempo da primeira inicialização não causar problemas de latência visíveis para sua aplicação, provavelmente não há necessidade de usar um pool ativo.

Tópicos

- Principais conceitos
- Pré-requisitos
- Atualização das instâncias em um pool ativo
- Recursos relacionados
- Limitações
- Usar ganchos do ciclo de vida com um grupo dinâmico no grupo do Auto Scaling
- Criar um pool ativo para seu grupo do Auto Scaling
- Visualizar o status e o motivo de falhas da verificação de integridade
- Exemplos para criar e gerenciar piscinas aquecidas com o AWS CLI

Principais conceitos

Antes de começar a usar, familiarize-se com os seguintes conceitos principais:

Pool ativo

Um pool ativo é um grupo de instâncias do EC2 pré-inicializadas que permanece ao lado de um grupo do Auto Scaling. Sempre que é necessário aumentar a escala da aplicação na horizontal, o grupo do Auto Scaling pode utilizar o pool ativo para atender à nova capacidade desejada. Isso o ajuda a garantir que as instâncias estejam prontas para começar rapidamente a servir o tráfego das aplicações, acelerando a resposta a um evento de aumento de escala na horizontal. Quando as instâncias deixam o pool ativo, elas passam a contar para a capacidade desejada do grupo. Isso é conhecido como inicialização a quente.

Enquanto as instâncias estão no pool quente, suas políticas de escalabilidade só são dimensionadas se o valor da métrica das instâncias que estão no estado `InService` for maior que o limite alto de alarme da política da escalabilidade (que é o mesmo que a utilização de destino de uma política de dimensionamento com monitoramento do objetivo).

Tamanho do pool ativo

Por padrão, o tamanho do pool ativo é calculado como a diferença entre a capacidade máxima do grupo do Auto Scaling e a capacidade desejada. Por exemplo, se a capacidade desejada do grupo do Auto Scaling for 6 e a capacidade máxima for 10, o tamanho do pool ativo será 4 quando você configurar o pool ativo pela primeira vez e o pool estiver inicializando.

Para especificar a capacidade máxima do grupo dinâmico separadamente, use a opção de especificação personalizada (`MaxGroupPreparedCapacity`) e defina um valor personalizado para ela que seja maior do que a capacidade atual do grupo. Se você fornecer um valor personalizado, o tamanho do grupo dinâmico será calculado como a diferença entre o valor personalizado e a capacidade desejada atual do grupo. Por exemplo, se a capacidade desejada do seu grupo do Auto Scaling for 6, se a capacidade máxima for 20 e se o valor personalizado for 8, o tamanho do seu grupo dinâmico será 2 quando você configurar o grupo dinâmico pela primeira vez e o grupo estiver sendo inicializado.

Talvez você só precise usar a opção de especificação personalizada (`MaxGroupPreparedCapacity`) ao trabalhar com grandes grupos do Auto Scaling para gerenciar os benefícios de custo de ter um grupo dinâmico. Por exemplo, talvez um grupo do Auto Scaling com 1.000 instâncias, uma capacidade máxima de 1.500 (para fornecer capacidade extra durante picos de tráfego de emergência) e um pool ativo de 100 instâncias seja uma estratégia melhor para ajudar você a atingir seus objetivos do que manter 500 instâncias reservadas para uso futuro no pool ativo.

Tamanho mínimo do pool ativo

Considere usar a configuração de tamanho mínimo (`MinSize`) para definir de modo estático o número mínimo de instâncias a serem mantidas no grupo de alta atividade. Não há tamanho mínimo definido por padrão. A configuração `MinSize` é útil quando você especifica `MaxGroupPreparedCapacity` para garantir que um número mínimo de instâncias seja mantido no grupo de alta atividade, mesmo quando a capacidade desejada do grupo do Auto Scaling for maior que `MaxGroupPreparedCapacity`.

Estado da instância do grupo de alta atividade

Você pode manter as instâncias no pool ativo em um de três estados: `Stopped`, `Running`, ou `Hibernated`. Manter as instâncias no estado `Stopped` é uma maneira eficaz de minimizar os custos. Com as instâncias interrompidas, você paga apenas pelos volumes usados e pelos endereços IP elásticos anexados às instâncias.

Você também pode manter as instâncias em um estado `Hibernated` para interromper instâncias sem excluir o conteúdo da memória (RAM). Quando uma instância é hibernada, isso sinaliza ao sistema operacional para salvar o conteúdo da RAM no volume raiz do Amazon EBS. Quando você inicia a instância novamente, o volume raiz é restaurado ao seu estado anterior, e o conteúdo da RAM é recarregado. Enquanto as instâncias estão em hibernação, você paga somente pelos volumes do EBS, incluindo armazenamento para o conteúdo da RAM e os endereços IP elásticos anexados às instâncias.

Também é possível manter instâncias em um estado `Running` no pool ativo, mas isso é altamente desaconselhável a fim de evitar a geração de cobranças desnecessárias. Quando as instâncias são interrompidas ou hibernadas, você economiza o custo das próprias instâncias. Você paga pelas instâncias somente quando elas são executadas.

Ganchos do ciclo de vida

Você usa ganchos do ciclo de vida para colocar instâncias em um estado de espera para poder executar ações personalizadas nas instâncias. Ações personalizadas são executadas à medida que as instâncias são iniciadas ou antes de serem terminadas.

Em uma configuração de pool quente, os ganchos do ciclo de vida atrasam a interrupção ou hibernação das instâncias e a colocação em serviço durante um evento de expansão até que concluam a inicialização. Se você adicionar um pool ativo ao seu grupo do Auto Scaling sem um gancho do ciclo de vida, as instâncias que demorarem muito para concluir a inicialização poderão ser interrompidas ou hibernadas e, em seguida, colocadas em serviço durante um evento de aumento de escala na horizontal antes de estarem prontas.

Política de reutilização de instâncias

Por padrão, o Amazon EC2 Auto Scaling termina suas instâncias quando seu grupo do Auto Scaling reduz a escala na horizontal. Em seguida, ele inicia novas instâncias no pool ativo para substituir as instâncias que foram terminadas.

Se desejar devolver instâncias ao pool ativo, você poderá especificar uma política de reutilização de instâncias. Isso permite reutilizar instâncias que já estão configuradas para atender ao tráfego de aplicações. Para garantir que seu pool ativo não seja excessivamente provisionado, o Amazon EC2 Auto Scaling pode terminar instâncias no pool ativo para reduzir seu tamanho quando for maior do que o necessário, com base em suas configurações. Ao terminar instâncias no pool ativo, ele usa a política de término padrão para escolher quais instâncias terminar primeiro.

> **Important**
>
> Se você desejar hibernar instâncias em redução de escala na horizontal e houver instâncias existentes no grupo do Auto Scaling, elas deverão atender aos requisitos de hibernação de instâncias. Caso contrário, quando as instâncias forem devolvidas ao pool ativo, elas recuarão para serem interrompidas em vez de serem hibernadas.

> **Note**
>
> No momento, só é possível especificar uma política de reutilização de instâncias usando a AWS CLI ou um SDK. Esse recurso não está disponível no console.

Pré-requisitos

Antes de criar um pool ativo para seu grupo do Auto Scaling, decida como você usará ganchos do ciclo de vida para inicializar novas instâncias com um estado inicial apropriado.

Para realizar ações personalizadas em instâncias enquanto elas estão em estado de espera devido a um gancho do ciclo de vida, você tem duas opções:

- Para cenários simples em que você deseja executar comandos em suas instâncias no início, você pode incluir um script de dados do usuário ao criar um modelo de execução ou configuração de execução para o grupo do Auto Scaling. Os scripts de dados do usuário são apenas scripts de shell normais ou diretivas do cloud-init init que são executadas pelo cloud-init quando as instâncias

são iniciadas. O script também pode controlar quando as instâncias fazem a transição para o próximo estado usando o ID da instância na qual é executado. Se você já não estiver fazendo isso, atualize seu script para recuperar o ID da instância nos metadados da instância. Para obter mais informações, consulte Acesso a metadados da instância no Guia do usuário do Amazon EC2.

> **Tip**
>
> Para executar scripts de dados do usuário quando uma instância é reiniciada, os dados do usuário devem estar no formato MIME de várias partes e especificar o seguinte na seção `#cloud-config` dos dados do usuário:
>
> ```
> #cloud-config
> cloud_final_modules:
>  - [scripts-user, always]
> ```

- Para cenários avançados em que você precisa de um serviço, como AWS Lambda fazer algo quando as instâncias entram ou saem do pool aquecido, você pode criar um gancho de ciclo de vida para seu grupo de Auto Scaling e configurar o serviço de destino para realizar ações personalizadas com base nas notificações do ciclo de vida. Para obter mais informações, consulte Destinos de notificação compatíveis.

Preparar instâncias para hibernação

Para preparar instâncias do Auto Scaling para usar o estado de grupo `Hibernated`, crie um novo modelo de execução ou configuração de execução configurada corretamente para oferecer suporte à hibernação de instância, conforme descrito no tópico Pré-requisitos de hibernação no Guia do usuário do Amazon EC2 para instâncias do Linux. Em seguida, associe o novo modelo de execução ou a configuração de execução ao grupo do Auto Scaling e inicie uma atualização de instância para substituir as instâncias associadas a um modelo de execução ou a uma configuração de execução anterior. Para obter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling.

Atualização das instâncias em um pool ativo

Para atualizar as instâncias em um pool ativo, você cria um novo modelo de execução ou configuração de execução e o associa ao grupo do Auto Scaling. Todas as novas instâncias serão iniciadas usando a nova AMI e outras atualizações especificadas no modelo de execução ou na configuração de execução, mas as instâncias existentes não serão afetadas.

Para forçar as instâncias do pool ativo de substituição a executar esse uso, o modelo de execução ou a configuração de execução, é possível iniciar uma atualização de instância para fazer uma atualização contínua de seu grupo. Uma atualização de instância substitui primeiro as instâncias `InService`. Em seguida, ela substitui as instâncias no pool ativo. Para obter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling.

Recursos relacionados

Você pode visitar nosso GitHubrepositório para ver exemplos de ganchos de ciclo de vida para piscinas aquecidas.

Limitações

- Limitações do grupo de alta atividade para um grupo do Auto Scaling com tipos de instâncias mistos:
  - Não há suporte aos grupos de alta atividade com grupos de instâncias mistas ponderadas. Se seu grupo do Auto Scaling usa ponderação de instâncias, não será possível adicionar um grupo de alta atividade.
  - Os grupos de alta atividade não oferecem suporte a instâncias spot em grupos de instâncias mistas. Sua política de instâncias mistas deve ser configurada para instâncias sob demanda somente ao usar grupos de alta atividade.
  - Ao usar grupos de alta atividade com grupos de instâncias mistas em estado de hibernação, será necessário configurar `HibernationOptions` em seu modelo de execução.
- O Amazon EC2 Auto Scaling pode colocar uma instância em um estado `Stopped` ou `Hibernated` somente quando ele tem um volume do Amazon EBS como dispositivo raiz. Instâncias que usam armazenamento de instâncias para o dispositivo raiz não podem ser interrompidas ou hibernadas.
- O Amazon EC2 Auto Scaling poderá colocar uma instância em um estado `Hibernated` somente se atender a todos os requisitos listados no tópico Pré-requisitos de hibernação no Guia do usuário do Amazon EC2 para instâncias do Linux.
- Se o pool ativo se esgotar em meio a um evento aumento de escala horizontal, as instâncias serão iniciadas diretamente no grupo do Auto Scaling (uma inicialização de baixa atividade). Uma inicialização de baixa atividade também poderá ocorrer se uma zona de disponibilidade estiver sem capacidade.
- Se uma instância dentro do grupo dinâmico encontrar um problema durante o processo de execução, impedindo-a de atingir o estado `InService`, a instância será considerada uma

execução defeituosa e será terminada. Isso se aplica independentemente da causa subjacente, como um erro de capacidade insuficiente ou qualquer outro fator.

- Se você tentar usar pools ativos com um grupo de nós gerenciados do Amazon Elastic Kubernetes Service (Amazon EKS), as instâncias que ainda estão sendo inicializadas poderão se registrar no cluster do Amazon EKS. Como resultado, o cluster pode agendar trabalhos em uma instância enquanto se prepara para ser interrompido ou hibernado.
- Da mesma forma, se você tentar usar um pool ativo com um cluster do Amazon ECS, as instâncias poderão se registrar no cluster antes que a inicialização seja concluída. Para resolver esse problema, é necessário configurar um modelo de inicialização ou uma configuração de inicialização que inclua uma variável de configuração de agente especial nos dados do usuário. Para obter mais informações, consulte Using a warm pool for your Auto Scaling group (Usar um grupo de alta atividade para o grupo do Auto Scaling) no Amazon Elastic Container Service Developer Guide (Guia do desenvolvedor do Amazon Elastic Container Service).

Usar ganchos do ciclo de vida com um grupo dinâmico no grupo do Auto Scaling

As instâncias em um pool ativo mantêm seu próprio ciclo de vida independente para ajudar você a criar a ação personalizada apropriada para cada transição. Esse ciclo de vida foi desenvolvido para ajudar você a invocar ações em um serviço-alvo (por exemplo, uma função do Lambda) enquanto uma instância ainda está sendo inicializada e antes de ser colocada em serviço.

> **Note**
>
> As operações de API que você usa para adicionar e gerenciar ganchos do ciclo de vida e concluir ações do ciclo de vida não são alteradas. Somente o ciclo de vida da instância é alterado.

Para obter mais informações sobre a adição de um gancho do ciclo de vida, consulte Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling. Para obter mais informações sobre a conclusão de uma ação do ciclo de vida, consulte Concluir uma ação do ciclo de vida em um grupo do Auto Scaling.

Para instâncias que entram no pool ativo, talvez você precise de um gancho do ciclo de vida por um dos seguintes motivos:

- Você deseja iniciar instâncias do EC2 via uma AMI que demora muito para concluir a inicialização.
- Você deseja executar scripts de dados do usuário para inicializar as instâncias do EC2.

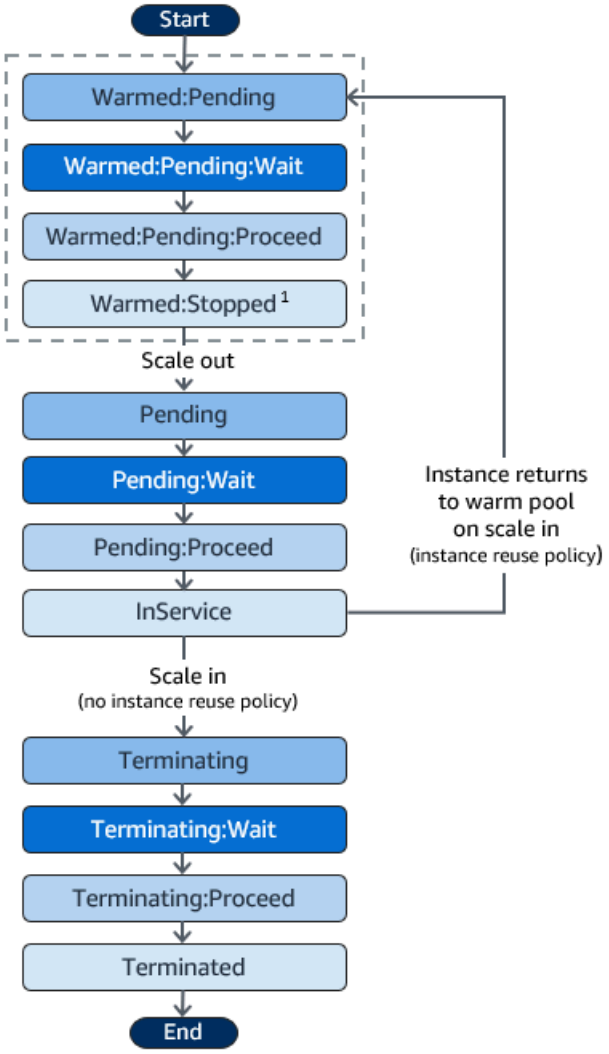
Para instâncias que saem do pool ativo, talvez você precise de um gancho do ciclo de vida por um dos seguintes motivos:

- Você pode usar algum tempo extra para preparar instâncias do EC2 para uso. Por exemplo, você pode ter serviços que devem ser iniciados quando uma instância é reiniciada antes que a aplicação possa funcionar corretamente.
- Você deseja preencher previamente os dados de cache para que um novo servidor não seja iniciado com um cache vazio.
- Você deseja registrar novas instâncias como instâncias gerenciadas com seu serviço de gerenciamento de configuração.

Transições de estado do ciclo de vida para instâncias em um pool ativo

Uma instância do Auto Scaling pode fazer a transição por muitos estados como parte de seu ciclo de vida.

O diagrama a seguir mostra a transição entre estados do Auto Scaling quando você usa um pool ativo:



[1] Esse estado varia de acordo com a configuração do estado do pool ativo. Se o estado do grupo estiver definido como `Running`, então esse estado será `Warmed:Running`, em vez disso. Se o estado do grupo estiver definido como `Hibernated`, então esse estado será `Warmed:Hibernated`, em vez disso.

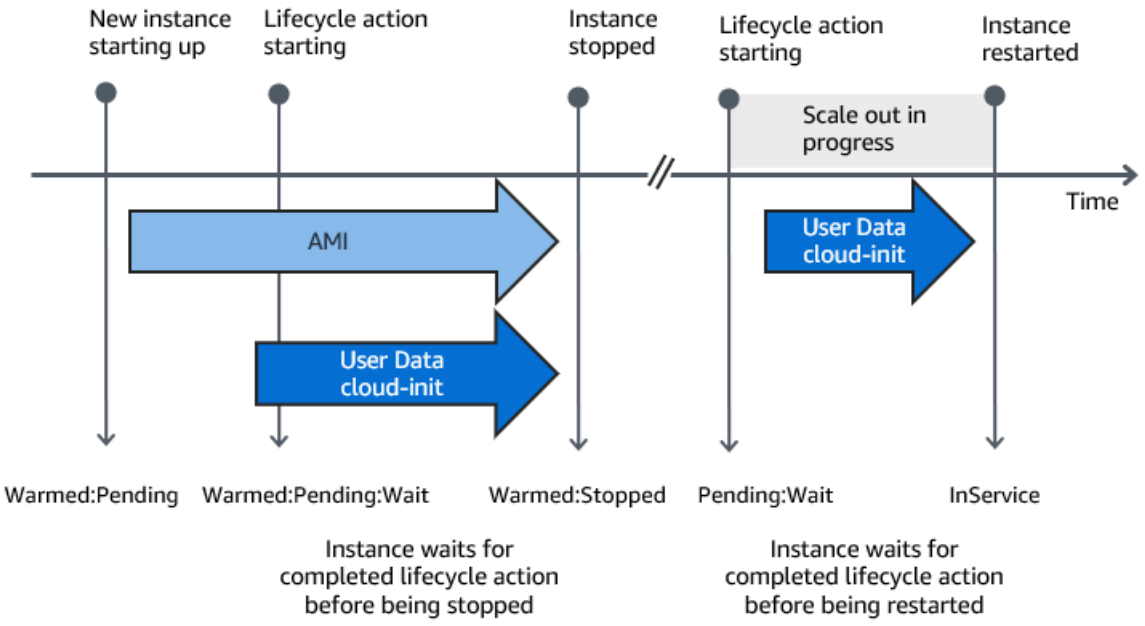
Ao adicionar ganchos do ciclo de vida, considere o seguinte:

- Quando um gancho do ciclo de vida é configurado para a ação `autoscaling:EC2_INSTANCE_LAUNCHING` do ciclo de vida, uma instância recém-iniciada faz uma primeira pausa para realizar uma ação personalizada quando atinge o estado `Warmed:Pending:Wait` e, novamente, quando a instância for reiniciada e atingir o estado `Pending:Wait`.
- Quando um gancho do ciclo de vida é configurado para a ação `EC2_INSTANCE_TERMINATING` do ciclo de vida, uma instância em encerramento faz uma pausa para realizar uma ação

personalizada quando atinge o estado `Terminating:Wait`. No entanto, se você especificar uma política de reutilização de instâncias para retornar instâncias ao pool ativo na operação de redução da escala horizontalmente em vez de encerrá-las, uma instância que estiver retornando ao pool ativo fará uma pausa para realizar uma ação personalizada no estado `Warmed:Pending:Wait` para a ação do ciclo de vida `EC2_INSTANCE_TERMINATING`.

- Se a demanda em sua aplicação esgotar o pool ativo, o Amazon EC2 Auto Scaling poderá iniciar instâncias diretamente no grupo do Auto Scaling se o grupo ainda não tiver atingido sua capacidade máxima. Se as instâncias forem executadas diretamente no grupo, elas só serão pausadas para realizar uma ação personalizada no estado `Pending:Wait`.
- Para controlar por quanto tempo uma instância permanece em um estado de espera antes de fazer a transição para o próximo estado, configure sua ação personalizada para usar o comando complete-lifecycle-action. Com os ganchos do ciclo de vida, as instâncias permanecem em estado de espera até que você notifique o Amazon EC2 Auto Scaling de que a ação especificada do ciclo de vida foi concluída, ou até que o período de tempo limite termine (uma hora, por padrão).

A seguir, um resumo do fluxo para um evento de aumento da escala na horizontal.



Quando as instâncias atingem um estado de espera, o Amazon EC2 Auto Scaling envia uma notificação. Exemplos dessas notificações estão disponíveis na EventBridge seção deste guia. Para obter mais informações, consulte Exemplos de eventos e padrões de pool ativo.

Destinos de notificação compatíveis

O Amazon EC2 Auto Scaling oferece suporte para definir qualquer um dos seguintes destinos como destinos de notificação para notificações do ciclo de vida:

- EventBridge regras
- Tópicos do Amazon SNS
- Filas do Amazon SQS
- AWS Lambda funções

> **Important**
>
> Se você tiver um script de dados do usuário (cloud-init) no modelo de execução ou na configuração de execução que configura as instâncias quando elas são iniciadas, você não precisará receber notificações para realizar ações personalizadas nas instâncias que estão sendo iniciadas ou reiniciadas.

As seções a seguir contêm links para a documentação que descreve como configurar destinos de notificação:

EventBridge regras — Para executar código quando o Amazon EC2 Auto Scaling coloca uma instância em estado de espera, você pode EventBridge criar uma regra e especificar uma função Lambda como destino. Para invocar diferentes funções do Lambda com base em notificações do ciclo de vida diferentes, você pode criar várias regras e associar cada regra a um padrão de evento específico e função do Lambda. Para obter mais informações, consulte Crie EventBridge regras para eventos em piscinas aquecidas.

Tópicos do Amazon SNS: para receber uma notificação quando uma instância é colocada em um estado de espera, você cria um tópico do Amazon SNS e, em seguida, configura a filtragem de mensagens do Amazon SNS para entregar notificações do ciclo de vida de forma diferente com base em um atributo de mensagem. Para obter mais informações, consulte Receba notificações usando o Amazon SNS.

Filas do Amazon SQS: para configurar um ponto de entrega para notificações do ciclo de vida em que um consumidor relevante possa buscá-las e processá-las, é possível criar uma fila do Amazon SQS e um consumidor de fila que processe mensagens da fila SQS. Se você quiser

que o consumidor da fila processe notificações do ciclo de vida de forma diferente com base em um atributo da mensagem, você também deverá configurar o consumidor da fila para analisar a mensagem e, em seguida, agir sobre a mensagem quando um atributo específico corresponder ao valor desejado. Para obter mais informações, consulte Receba notificações usando o Amazon SQS.

AWS Lambda funções — Para executar código personalizado quando o Amazon EC2 Auto Scaling coloca uma instância em estado de espera, você pode especificar uma função Lambda como destino da notificação. A função Lambda é invocada com dados de notificação do ciclo de vida, permitindo que você execute ações personalizadas, como configuração de instância, configuração de aplicativos ou integração com outros serviços. AWS É necessário configurar a política baseada em recursos da função do Lambda para permitir que o perfil vinculado ao serviço do Auto Scaling invoque a função. Para obter mais informações, consulte Encaminhe notificações AWS Lambda diretamente para.

Criar um pool ativo para seu grupo do Auto Scaling

Este tópico descreve como criar um pool ativo para seu grupo do Auto Scaling.

> **Important**
>
> Antes de continuar, preencha os pré-requisitos para criar um pool ativo e confirme se você criou um gancho do ciclo de vida para o grupo do Auto Scaling.

Criar um pool ativo

Use o procedimento a seguir para criar um pool ativo para o grupo do Auto Scaling.

Para criar um pool ativo (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado de um grupo existente.

   Um painel dividido é aberto na parte inferior da página.
3. Selecione a guia Instance management (Gerenciamento de instâncias).
4. Em Warm pool (pool ativo), escolha Create warm pool (Criar pool ativo).
5. Para configurar um pool ativo, faça o seguinte:

a. Em Warm pool instance state (Estado da instância do pool de alta atividade), escolha para qual estado você deseja fazer a transição das instâncias quando elas entrarem no pool ativo. O padrão é `Stopped`.

b. Em Minimum warm pool size (Tamanho mínimo do pool ativo), insira o número mínimo de instâncias que serão mantidas no pool ativo.

c. Para a reutilização de instâncias, marque a caixa de seleção Reutilizar em redução da escala horizontalmente para permitir que as instâncias no grupo do Auto Scaling retornem ao grupo dinâmico em redução da escala horizontalmente.

d. Para o Tamanho do grupo dinâmico, escolha uma das opções disponíveis:

- Especificação padrão: O tamanho do grupo dinâmico é determinado pela diferença entre a capacidade máxima e a desejada do grupo do Auto Scaling. Essa opção simplifica o gerenciamento do grupo dinâmico. Depois de criar o grupo dinâmico, seu tamanho pode ser facilmente atualizado, bastando ajustar a capacidade máxima do grupo.
- Especificação personalizada: O tamanho do grupo dinâmico é determinado pela diferença entre um valor personalizado e a capacidade desejada do grupo do Auto Scaling. Essa opção lhe dá flexibilidade para gerenciar o tamanho de seu grupo dinâmico independentemente da capacidade máxima do grupo.

6. Consulte a seção Tamanho estimado do grupo dinâmico com base nas configurações atuais para confirmar como a especificação padrão ou personalizada se aplica ao tamanho do esse grupo. Lembre-se de que o tamanho do grupo dinâmico depende da capacidade desejada do grupo do Auto Scaling, que será alterada se o grupo for escalado.
7. Escolha Criar.

Seleção do tipo de instância com grupos de instâncias mistas

O Auto Scaling prioriza os tipos de instância que já estejam no grupo de alta atividade durante eventos de escalabilidade quando seu grupo está configurado com uma política de instâncias mistas. Comportamento da execução:

1. O Auto Scaling tenta iniciar instâncias usando os tipos de instância disponíveis no grupo de alta atividade.
2. Se a inicialização a quente falhar, o Auto Scaling tentará a inicialização a frio usando todos os tipos de instâncias restantes em sua política de instâncias mistas.

Example

Exemplo

Se você configurar seu grupo do Auto Scaling com 10 tipos de instâncias e seu grupo de alta atividade contiver 6 desses tipos de instâncias. Durante o aumento da escala na horizontal, o Auto Scaling primeiro experimenta os 6 tipos de instâncias do grupo de alta atividade. Se não obtiver êxito, o Auto Scaling tentará todos os tipos de instância configurados por meio de inicialização a frio.

Isso oferece benefícios de desempenho de grupo de alta atividade sempre que possível, mantendo a flexibilidade de sua configuração completa de instâncias mistas.

Excluir um grupo de alta atividade

Quando você não precisar mais do pool ativo, use o procedimento a seguir para excluí-lo.

Para excluir o grupo de alta atividade (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado de um grupo existente.

   Um painel dividido é aberto na parte inferior da página.
3. Selecione a guia Instance management (Gerenciamento de instâncias).
4. Em Warm pool (pool ativo), escolha Actions (Ações), Delete (Excluir).
5. Quando a confirmação for solicitada, escolha Excluir.

Visualizar o status e o motivo de falhas da verificação de integridade

As verificações de integridade permitem que o Amazon EC2 Auto Scaling determine quando uma instância não está íntegra e deve ser terminada. Para instâncias de pool ativo mantidas em um estado `Stopped`, ele emprega o conhecimento que o Amazon EBS tem da disponibilidade de uma instância `Stopped` para identificar instâncias não íntegras. Ele faz isso chamando a API `DescribeVolumeStatus` para determinar o status do volume do EBS anexado à instância. Para instâncias de pool ativo mantidas em um estado `Running`, ele depende das verificações de status do EC2 para determinar a integridade da instância. Embora não haja período de carência de verificação de integridade para instâncias de pools ativos, o Amazon EC2 Auto Scaling não começará a verificar a integridade da instância até que o gancho do ciclo de vida seja concluído.

Quando uma instância não está íntegra, o Amazon EC2 Auto Scaling a exclui automaticamente e cria uma nova instância para substituí-la. Geralmente, as instâncias são terminadas dentro de alguns minutos após a falha na verificação de integridade. Para obter mais informações, consulte Veja o motivo das falhas na verificação de integridade.

Verificações de integridade personalizadas também são aceitas. Isso poderá ser útil se você tiver seu próprio sistema de verificação de integridade capaz de detectar a integridade de uma instância e enviar essas informações para o Amazon EC2 Auto Scaling. Para obter mais informações, consulte Configuração de uma verificação de integridade personalizada para seu grupo do Auto Scaling.

No console do Amazon EC2 Auto Scaling, é possível visualizar o status (íntegra ou não íntegra) das instâncias do pool ativo. Você também pode ver o estado de saúde deles usando o AWS CLI ou um dos SDKs.

Para visualizar o status das instâncias do grupo de alta atividade (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Instance management (Gerenciamento de instâncias), em Warm pool instances (Instâncias do pool ativo), a coluna Lifecycle (Ciclo de vida) contém o estado das instâncias.

   A coluna Health status (Status da integridade) mostra a avaliação da integridade da instância feita pelo Amazon EC2 Auto Scaling.

   > **Note**
   >
   > As novas instâncias começam íntegras. Até que o gancho do ciclo de vida seja concluído, a integridade de uma instância não será verificada.

Para visualizar o motivo das falhas de verificação de integridade (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).

3. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status mostra se o seu grupo do Auto Scaling iniciou ou terminou instâncias com êxito.

   Se ele terminou quaisquer instâncias não íntegras, a coluna Cause (Causa) mostrará a data e a hora do término e o motivo da falha na verificação de integridade. Por exemplo, "At 2021-04-01T21:48:35Z an instance was taken out of service in response to EBS volume health check failure" (Em 2021-04-01T 21:48:35 Z uma instância foi retirada de serviço em resposta a falha na verificação de integridade do volume do EBS).

Para visualizar o status das instâncias do pool ativo (AWS CLI)

Visualize a piscina aquecida de um grupo de Auto Scaling usando o comando a seguir describe-warm-pool.

```
aws autoscaling describe-warm-pool --auto-scaling-group-name my-asg
```

Saída de exemplo.

```
{
    "WarmPoolConfiguration": {
        "MinSize": 0,
        "PoolState": "Stopped"
    },
    "Instances": [
        {
            "InstanceId": "i-0b5e5e7521cfaa46c",
            "InstanceType": "t2.micro",
            "AvailabilityZone": "us-west-2a",
            "LifecycleState": "Warmed:Stopped",
            "HealthStatus": "Healthy",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-08c4cd42f320d5dcd",
                "LaunchTemplateName": "my-template-for-auto-scaling",
                "Version": "1"
            }
        },
        {
            "InstanceId": "i-0e21af9dcfb7aa6bf",
```

```
            "InstanceType": "t2.micro",
            "AvailabilityZone": "us-west-2a",
            "LifecycleState": "Warmed:Stopped",
            "HealthStatus": "Healthy",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-08c4cd42f320d5dcd",
                "LaunchTemplateName": "my-template-for-auto-scaling",
                "Version": "1"
            }
        }
    ]
}
```

Para visualizar o motivo das falhas de verificação de integridade (AWS CLI)

Use o seguinte comando describe-scaling-activities:

```
aws autoscaling describe-scaling-activities --auto-scaling-group-name my-asg
```

Esta é uma resposta de exemplo, em que `Description` indica que seu grupo do Auto Scaling encerrou uma instância e `Cause` indica o motivo da falha na verificação de integridade.

As ações de escalabilidade são ordenadas por horário de início. As atividades ainda em andamento são descritas primeiro.

```
{
  "Activities": [
    {
      "ActivityId": "4c65e23d-a35a-4e7d-b6e4-2eaa8753dc12",
      "AutoScalingGroupName": "my-asg",
      "Description": "Terminating EC2 instance: i-04925c838b6438f14",
      "Cause": "At 2021-04-01T21:48:35Z an instance was taken out of service in
 response to EBS volume health check failure.",
      "StartTime": "2021-04-01T21:48:35.859Z",
      "EndTime": "2021-04-01T21:49:18Z",
      "StatusCode": "Successful",
      "Progress": 100,
      "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-west-2a
\"...}",
      "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
```

```
    },
...
  ]
}
```

Exemplos para criar e gerenciar piscinas aquecidas com o AWS CLI

Você pode criar e gerenciar piscinas aquecidas usando o Console de gerenciamento da AWS, AWS Command Line Interface (AWS CLI) ou SDKs.

Os exemplos a seguir mostram como criar e gerenciar pools ativos usando a AWS CLI.

Conteúdo

- Exemplo 1: manter instâncias no estado Stopped
- Exemplo 2: manter instâncias no estado Running
- Exemplo 3: manter instâncias no estado Hibernated
- Exemplo 4: retornar instâncias para o pool ativo ao reduzir a escala na horizontal
- Exemplo 5: especificar o número mínimo de instâncias no pool ativo
- Exemplo 6: Definir o tamanho do grupo dinâmico usando uma especificação personalizada.
- Exemplo 7: definir um tamanho de pool ativo absoluto
- Exemplo 8: exclusão um pool ativo

Exemplo 1: manter instâncias no estado **`Stopped`**

O put-warm-poolexemplo a seguir cria um pool aquecido que mantém as instâncias em um `Stopped` estado.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped
```

Exemplo 2: manter instâncias no estado **`Running`**

O put-warm-poolexemplo a seguir cria um pool aquecido que mantém as instâncias em um `Running` estado em vez de em um `Stopped` estado.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
```

```
  --pool-state Running
```

Exemplo 3: manter instâncias no estado **`Hibernated`**

O put-warm-poolexemplo a seguir cria um pool aquecido que mantém as instâncias em um `Hibernated` estado em vez de em um `Stopped` estado. Isso permite interromper instâncias sem excluir o conteúdo da memória (RAM).

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Hibernated
```

Exemplo 4: retornar instâncias para o pool ativo ao reduzir a escala na horizontal

O put-warm-poolexemplo a seguir cria um pool aquecido que mantém as instâncias em um `Stopped` estado e inclui a `--instance-reuse-policy` opção. O valor da política de reutilização de instâncias `'{"ReuseOnScaleIn": true}'` informa ao Amazon EC2 Auto Scaling para devolver as instâncias ao pool ativo quando o grupo do Auto Scaling reduz a escala na horizontal.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --instance-reuse-policy '{"ReuseOnScaleIn": true}'
```

Exemplo 5: especificar o número mínimo de instâncias no pool ativo

O put-warm-poolexemplo a seguir cria um pool aquecido que mantém no mínimo 4 instâncias, para que haja pelo menos 4 instâncias disponíveis para lidar com picos de tráfego.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --min-size 4
```

Exemplo 6: Definir o tamanho do grupo dinâmico usando uma especificação personalizada.

Por padrão, o Amazon EC2 Auto Scaling gerencia o tamanho do seu grupo dinâmico como a diferença entre a capacidade máxima e a desejada do grupo do Auto Scaling. No entanto, você pode gerenciar o tamanho do grupo dinâmico independentemente da sua capacidade máxima, usando a opção `--max-group-prepared-capacity`.

O put-warm-poolexemplo a seguir cria um pool aquecido e define o número máximo de instâncias que podem existir simultaneamente no pool aquecido e no grupo Auto Scaling. Se o grupo tiver

uma capacidade desejada de 800, o grupo dinâmico terá inicialmente um tamanho de 100 ao ser inicializado após a execução desse comando.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --max-group-prepared-capacity 900
```

Para manter um número mínimo de instâncias no pool ativo, inclua a opção `--min-size` com o comando, da seguinte forma.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --max-group-prepared-capacity 900 --min-size 25
```

Exemplo 7: definir um tamanho de pool ativo absoluto

Se você definir os mesmos valores para as opções `--max-group-prepared-capacity` e `--min-size`, o pool ativo terá um tamanho absoluto. O put-warm-poolexemplo a seguir cria um pool quente que mantém um tamanho constante de pool quente de 10 instâncias.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --min-size 10 --max-group-prepared-capacity 10
```

Exemplo 8: exclusão um pool ativo

Use o delete-warm-poolcomando a seguir para excluir uma piscina aquecida.

```
aws autoscaling delete-warm-pool --auto-scaling-group-name my-asg
```

Se houver instâncias no pool aquecido ou se atividades de escalonamento estiverem em andamento, use o delete-warm-poolcomando com a `--force-delete` opção. Essa opção também terminará as instâncias do Amazon EC2 e quaisquer ações do ciclo de vida pendentes.

```
aws autoscaling delete-warm-pool --auto-scaling-group-name my-asg --force-delete
```

Deslocamento zonal do grupo Auto Scaling

A mudança zonal é um recurso do Amazon Application Recovery Controller (ARC). Com a mudança zonal, você pode se recuperar rapidamente das deficiências do aplicativo em uma zona de

disponibilidade com uma única ação. Quando você ativa o deslocamento zonal para um grupo do Auto Scaling, o grupo é registrado no serviço de mudança zonal ARC. Em seguida, você pode iniciar uma mudança de zona usando a API Console de gerenciamento da AWS AWS CLI, ou e o grupo Auto Scaling trata a zona de disponibilidade com uma mudança de zona ativa como prejudicada.

Conceitos de mudança zonal de grupos do Auto Scaling

Antes de continuar, certifique-se de estar familiarizado com os seguintes conceitos básicos relacionados à integração com o deslocamento zonal ARC.

Mudança zonal ARC

O Auto Scaling pode registrar grupos de Auto Scaling com deslocamento zonal ARC quando você ativa esse recurso. Após o registro, você pode visualizar seus recursos com a `ListManagedResources` API ARC. Para obter mais informações, consulte Mudança zonal no ARC no Guia do desenvolvedor do Amazon Application Recovery Controller (ARC).

Rebalanceamento de zonas de disponibilidade

O Auto Scaling tenta manter a capacidade equilibrada em cada zona de disponibilidade. Quando ocorre um desequilíbrio entre as zonas de disponibilidade, o Auto Scaling tenta corrigir automaticamente o desequilíbrio. Para obter mais informações, consulte Distribuição de instâncias.

Escalabilidade dinâmica

O escalonamento dinâmico dimensiona a capacidade desejada do seu grupo de Auto Scaling com base nas métricas que você escolhe com as políticas de escalabilidade. Para obter mais informações, consulte Escalabilidade dinâmica para o Amazon EC2 Auto Scaling.

Verificações de integridade

O Auto Scaling verifica periodicamente o status de integridade de todas as instâncias dentro de um grupo de Auto Scaling para garantir que elas estejam funcionando e em boas condições. Quando uma instância não íntegra é detectada, o Auto Scaling a marca para substituição. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Atualização de instância

Você inicia as atualizações de instâncias para atualizar as instâncias em seu grupo do Auto Scaling. Depois que uma atualização de instância é iniciada, o Auto Scaling tenta substituir todas

as instâncias em seu grupo de Auto Scaling. Para obter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling.

Pré-escalado

Você pode tolerar a perda de uma única zona de disponibilidade porque tem capacidade suficiente nas zonas de disponibilidade restantes para seu aplicativo.

Aumento de escala

Quando você aumenta a capacidade desejada de um grupo de Auto Scaling, o Auto Scaling tenta iniciar instâncias adicionais para atender à nova capacidade desejada. Por padrão, o Auto Scaling inicia a instância de forma equilibrada para manter a mesma capacidade em cada zona de disponibilidade ativada em um grupo de Auto Scaling.

Como a mudança zonal funciona para grupos de Auto Scaling

Suponha que você tenha um grupo de Auto Scaling com as seguintes zonas de disponibilidade:

- `us-east-1a`
- `us-east-1b`
- `us-east-1c`

Você ativou a mudança zonal em todas as zonas de disponibilidade e percebe falhas`us-east-1a`, então você aciona uma mudança zonal. Os comportamentos a seguir ocorrem quando uma mudança zonal é acionada`us-east-1a`.

- Escalabilidade horizontal — O Auto Scaling lançará todas as novas solicitações de capacidade nas zonas de disponibilidade saudáveis `us-east-1b` (`us-east-1c`e).
- Escalabilidade dinâmica — O Auto Scaling impedirá que as políticas de escalabilidade diminuam a capacidade desejada em todas as zonas de disponibilidade. O Auto Scaling não impedirá que as políticas de escalabilidade aumentem a capacidade desejada em todas as zonas de disponibilidade.
- Atualizações de instância — O Auto Scaling estenderá o tempo limite para qualquer processo de atualização de instância que esteja atrasado enquanto uma mudança de zona estiver ativa.

A tabela a seguir descreve o comportamento da verificação de integridade de cada opção quando uma mudança de zona é acionada`us-east-1a`.

Seleção de comportamento de verificação de integrida de da zona de disponibilidade prejudicada	Comportamento de verificação de saúde			
Substitua insalubre	As instâncias que parecem insalubres serão substituídas em todas as zonas de disponibi lidade (`us-east-1a us-east-1b` , `eus-east-1 c`).			
Ignore os não saudáveis	As instâncias que parecem insalubres serão substituídas em `us-east-1b` e. `us-east-1 c` As instância s não serão substituídas na zona de disponibilidade pelo deslocame nto zonal ativo (`us-east-1a`).			

Melhores práticas para usar o deslocamento zonal

Para manter a alta disponibilidade de seus aplicativos ao usar o deslocamento zonal, recomendamos as seguintes melhores práticas:

- Monitore EventBridge as notificações para determinar quando há um evento contínuo de comprometimento da Zona de Disponibilidade. Para obter mais informações, consulte Use EventBridge para lidar com eventos do Auto Scaling.
- Use políticas de escalabilidade com limites apropriados para garantir que você tenha capacidade suficiente para tolerar a perda de uma zona de disponibilidade.
- Defina uma política de manutenção de instâncias com uma porcentagem íntegra mínima de 100. Com essa configuração, o Auto Scaling espera que uma nova instância esteja pronta para uso antes de encerrar uma instância não íntegra.

Para clientes pré-escalados, também recomendamos o seguinte:

- Selecione Ignorar não íntegro como o comportamento de verificação de integridade da Zona de Disponibilidade comprometida, pois você não precisa substituir a instância não íntegra durante o evento de comprometimento.
- Use o deslocamento automático zonal no ARC para seus grupos de Auto Scaling. O recurso de mudança automática zonal no ARC permite AWS deslocar o tráfego de um recurso para fora de uma zona de disponibilidade ao AWS detectar uma deficiência em uma zona de disponibilidade. Para obter mais informações, consulte Mudança automática zonal no ARC no Guia do desenvolvedor do Amazon Application Recovery Controller (ARC).

Para clientes com balanceadores de carga desativados em várias zonas, também recomendamos o seguinte:

- Use balanceado somente para sua distribuição na zona de disponibilidade.
- Se você estiver usando o deslocamento zonal nos grupos do Auto Scaling e nos balanceadores de carga, cancele primeiro o deslocamento zonal no seu grupo do Auto Scaling. Em seguida, aguarde o equilíbrio da capacidade em todas as zonas de disponibilidade antes de cancelar a mudança zonal no balanceador de carga.
- Devido à possibilidade de desequilíbrio de capacidade quando você ativa a mudança de zona e usa um balanceador de carga desativado entre zonas, o Auto Scaling inclui uma etapa de validação extra. Se você estiver seguindo as melhores práticas,

você pode reconhecer essa possibilidade marcando a Console de gerenciamento da AWS caixa de seleção ou usando a `skip-zonal-shift-validation` bandeira em`CreateAutoScalingGroup`,`UpdateAutoScalingGroup`, ou`AttachTrafficSources`.

Para obter mais informações sobre o uso do deslocamento zonal com grupos do Auto Scaling, consulte AWS o blog de computação Usando o deslocamento zonal com o Amazon EC2 Auto Scaling.

Ative o deslocamento zonal usando o Console de gerenciamento da AWS ou o AWS CLI

Para ativar o deslocamento zonal, use um dos métodos a seguir.

Console

Para habilitar a mudança zonal em um novo grupo (console)

1. Siga as instruções Criar um grupo do Auto Scaling usando um modelo de execução e conclua cada etapa do procedimento, até a etapa 10.
2. Na página Integrar com outros serviços, para o deslocamento zonal do Application Recovery Controller (ARC), marque a caixa de seleção para ativar o deslocamento zonal.
3. Em Comportamento de verificação de integridade, escolha Ignorar não íntegro ou Substituir não íntegro. Para obter mais informações, consulte Como a mudança zonal funciona para grupos de Auto Scaling.
4. Continue com as etapas em Criar um grupo do Auto Scaling usando um modelo de execução.

AWS CLI

Para habilitar a mudança zonal em um novo grupo ()AWS CLI

Adicione o parâmetro `--availability-zone-impairment-policy` ao comando create-auto-scaling-group.

O `--availability-zone-impairment-policy` parâmetro tem duas opções:

- ZonalShiftEnabled— Se definido como`true`, o Auto Scaling registra o grupo Auto Scaling com o deslocamento zonal ARC e você pode iniciar, atualizar ou cancelar um deslocamento zonal

no console ARC. Se definido como`false`, o Auto Scaling cancela o registro do grupo Auto Scaling do deslocamento zonal ARC. Você já deve ter a mudança de zona ativada para `false` definir como.

- ImpairedZoneHealthCheckBehavior— Se definido como`replace-unhealthy`, as instâncias não íntegras serão substituídas na Zona de Disponibilidade pela mudança de zona ativa. Se definido como`ignore-unhealthy`, as instâncias não íntegras não serão substituídas na Zona de Disponibilidade pela mudança de zona ativa. Para obter mais informações, consulte Como a mudança zonal funciona para grupos de Auto Scaling.

O exemplo a seguir permite a mudança de zona em um novo grupo de Auto Scaling chamado. *my-asg*

```
aws autoscaling create-auto-scaling-group \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --auto-scaling-group-name my-asg \
  --min-size 1 \
  --max-size 10 \
  --desired-capacity 5 \
  --availability-zones us-east-1a us-east-1b us-east-1c \
  --availability-zone-impairment-policy '{
      "ZonalShiftEnabled": true,
      "ImpairedZoneHealthCheckBehavior": IgnoreUnhealthy
    }'
```

Console

Para habilitar a mudança zonal em um grupo existente (console)

1. Abra o EC2 console da Amazon em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a mesma Região da AWS na qual você criou o grupo do Auto Scaling.
3. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
4. Na guia Integrações, em Deslocamento zonal do Controlador de Recuperação de Aplicativos (ARC), escolha Editar.

5. Marque a caixa de seleção para ativar a mudança de zona.
6. Em Comportamento de verificação de integridade, escolha Ignorar não íntegro ou Substituir não íntegro. Para obter mais informações, consulte Como a mudança zonal funciona para grupos de Auto Scaling.
7. Selecione Atualizar.

AWS CLI

Para habilitar a mudança zonal em um grupo existente ()AWS CLI

Adicione o parâmetro `--availability-zone-impairment-policy` ao comando update-auto-scaling-group.

O `--availability-zone-impairment-policy` parâmetro tem duas opções:

- ZonalShiftEnabled— Se definido como`true`, o Auto Scaling registra o grupo Auto Scaling com o deslocamento zonal ARC e você pode iniciar, atualizar ou cancelar um deslocamento zonal no console ARC. Se definido como`false`, o Auto Scaling cancela o registro do grupo Auto Scaling do deslocamento zonal ARC. Você já deve ter a mudança de zona ativada para `false` definir como.
- ImpairedZoneHealthCheckBehavior— Se definido como`replace-unhealthy`, as instâncias não íntegras serão substituídas na Zona de Disponibilidade pela mudança de zona ativa. Se definido como`ignore-unhealthy`, as instâncias não íntegras não serão substituídas na Zona de Disponibilidade pela mudança de zona ativa. Para obter mais informações, consulte Como a mudança zonal funciona para grupos de Auto Scaling.

O exemplo a seguir permite a mudança de zona no grupo de Auto Scaling especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --availability-zone-impairment-policy '{
      "ZonalShiftEnabled": true,
      "ImpairedZoneHealthCheckBehavior": IgnoreUnhealthy
    }'
```

Distribuição de zonas de disponibilidade de grupo do Auto Scaling

As informações a seguir descrevem as estratégias de zona de disponibilidade de grupo do Auto Scaling.

Melhor esforço balanceado

> O Auto Scaling mantém um número igual de instâncias nas zonas de disponibilidade habilitadas. Se as tentativas de execução falharem em uma zona de disponibilidade, o Auto Scaling tentará iniciar instâncias em outra zona de disponibilidade íntegra. Essa estratégia é importante para aplicações que precisem de redundância de zona de disponibilidade e não sejam afetadas por grupos desbalanceados.

Somente balanceado

> O Auto Scaling mantém um número igual de instâncias nas zonas de disponibilidade habilitadas. Se as tentativas de execução falharem em uma zona de disponibilidade, o Auto Scaling continuará tentando iniciar instâncias na zona de disponibilidade. Essa estratégia é importante para atender a determinados requisitos, como workloads baseadas em quórum ou se seu grupo do Auto Scaling puder tolerar a perda de uma zona de disponibilidade por você ter capacidade suficiente nas demais zonas de disponibilidade.

A seleção da estratégia de distribuição da Zona de Disponibilidade está na seção Rede do Console de gerenciamento da AWS ou você pode usar os update-auto-scaling-groupcomandos create-auto-scaling-groupou.

Para obter mais informações, consulte Criar grupos do Auto Scaling usando modelos de execução.

Desvincular instâncias do seu grupo do Auto Scaling

Você pode desvincular instâncias do seu grupo do Auto Scaling. Depois que uma instância é desanexada, ela se torna independente e pode ser gerenciada por conta própria ou anexada a um grupo do Auto Scaling diferente, separado do grupo original ao qual pertencia. Isso pode ser útil, por exemplo, quando você deseja realizar testes usando instâncias existentes que já estão executando a aplicação.

Este tópico fornece instruções sobre como desanexar e anexar instâncias. Ao anexar instâncias, você também pode usar uma instância existente em vez de uma desanexada.

Em vez de desanexar e reanexar uma instância ao mesmo grupo, recomendamos usar o procedimento de espera para remover temporariamente a instância do grupo. Para obter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.

Conteúdo

- Considerações para desanexar instâncias
- Considerações para anexar instâncias
- Mover uma instância para um grupo diferente usando desanexar e anexar

Considerações para desanexar instâncias

Ao desanexar instâncias, tenha em mente estes pontos:

- Será possível desvincular uma instância somente quando ela estiver nos estados `InService` ou `StandBy`. Se você estiver desanexando instâncias que estão no estado `StandBy`, adote cautela. Incluir o sinalizador `ShouldDecrementDesiredCapacity` na chamada de API ao tentar desvincular instâncias após colocá-las no estado `StandBy` pode fazer com que outras instâncias sejam encerradas inesperadamente.
- Depois que você desanexar uma instância, ela continuará em execução e incorrendo em cobranças. Para evitar cobranças desnecessárias, certifique-se de reanexar ou encerrar as instâncias desanexadas quando elas não forem mais necessárias.
- Você pode optar por diminuir a capacidade desejada pelo número de instâncias que está desanexando. Se você optar por não reduzir a capacidade, o Amazon EC2 Auto Scaling iniciará novas instâncias para substituir as instâncias desanexadas, mantendo a capacidade desejada.
- Se o número de instâncias que você está desanexando reduzirá o grupo de Auto Scaling abaixo de sua capacidade mínima, é necessário diminuir a capacidade mínima.
- Se você desanexar várias instâncias da mesma Zona de Disponibilidade sem diminuir a capacidade desejada, o grupo se reequilibrará automaticamente, a menos que você suspenda o processo da `AZRebalance`. Para obter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.
- Se você desvincular uma instância de um grupo do Auto Scaling que tenha um grupo de destino de balanceador de carga ou um Classic Load Balancer anexado, a instância será cancelada no balanceador de carga. Se a drenagem da conexão (atraso no cancelamento do registro) estiver habilitada para seu balanceador de carga, o Amazon EC2 Auto Scaling aguardará a conclusão das solicitações em andamento.

Considerações para anexar instâncias

Observe o seguinte ao anexar instâncias:

- O Amazon EC2 Auto Scaling trata as instâncias anexadas da mesma forma que as instâncias iniciadas pelo próprio grupo. Isso significa que as instâncias anexadas podem ser encerradas durante eventos de redução da escala horizontalmente se forem selecionadas. As permissões concedidas pelo perfil vinculado ao serviço AWSServiceRoleForAutoScaling permitem que o Amazon EC2 Auto Scaling faça isso.
- Quando você anexa instâncias, a capacidade desejada do grupo aumente de acordo com o número de instâncias que estão sendo anexadas. Se a capacidade desejada após adicionar novas instâncias exceder o tamanho máximo do grupo, a solicitação para anexar mais instâncias falhará.
- Se você adicionar instâncias ao seu grupo causando distribuição desigual entre zonas de disponibilidade, o Amazon EC2 Auto Scaling reequilibra o grupo para restabelecer uma distribuição uniforme, a menos que você suspenda o processo`AZRebalance`. Para obter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.
- Se você anexar uma instância a um grupo do Auto Scaling que tenha um grupo de destino de balanceador de carga ou um Classic Load Balancer anexado, a instância será registrada no balanceador de carga.

Para que uma instância seja anexada, ela deve atender aos seguintes critérios:

- A instância está no estado `running` com o Amazon EC2.
- A AMI usada para ativar a instância ainda deve existir.
- A instância não é um membro de outro grupo do Auto Scaling.
- A instância é executada em uma das zonas de disponibilidade definidas no grupo do Auto Scaling.
- Se o grupo do Auto Scaling tiver um grupo de destino de balanceador de carga ou Classic Load Balancer anexado, a instância e o balanceador de carga deverão estar ambos na mesma VPC.

Mover uma instância para um grupo diferente usando desanexar e anexar

Use um dos procedimentos a seguir para desanexar uma instância do seu grupo do Auto Scaling e anexá-la a um grupo diferente do Auto Scaling.

Para criar um novo grupo do Auto Scaling a partir de uma instância desanexada, consulte Crie um grupo de Auto Scaling a partir de uma instância existente usando o AWS CLI (não recomendado, cria uma configuração de inicialização).

Console

Para desvincular uma instância de um grupo do Auto Scaling existente

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), selecione uma instância e escolha Actions (Ações) e Detach (Desvincular).
4. Na caixa de diálogo Desanexar instância, mantenha a caixa de seleção Substituir instância marcada para iniciar uma instância substituta. Desmarque a caixa de seleção para diminuir a capacidade desejada.
5. Quando a confirmação for solicitada, digite **`detach`** para confirmar a exclusão da instância especificada do grupo do Auto Scaling e, em seguida, escolha Desvincular instância.

Agora você pode anexar a instância a um grupo do Auto Scaling diferente.

Como anexar uma instância a um grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. (Opcional) No painel de navegação, em Auto Scaling, escolha Grupos do Auto Scaling. Selecione o grupo do Auto Scaling e verifique se o tamanho máximo do grupo do Auto Scaling é grande o suficiente para que você possa adicionar outra instância. Caso contrário, na guia Detalhes aumente a capacidade máxima.
3. No painel de navegação, em Instances (Instâncias), escolha Instances (Instâncias) e selecione uma instância.
4. Escolha Actions (Ações), Instance settings (Configurações da instância), Attach to Auto Scaling Group (Anexar ao grupo do Auto Scaling).
5. Na página Attach to Auto Scaling group (Anexar ao grupo do Auto Scaling), em Auto Scaling Group (Grupo do Auto Scaling), selecione o grupo do Auto Scaling e, em seguida, escolha Attach (Anexar).

6. Se a instância não atender aos critérios, você receberá uma mensagem de erro com os detalhes. Por exemplo, a instância pode não estar na mesma zona de disponibilidade que o grupo do Auto Scaling. Escolha Close (Fechar) e tente novamente com um Grupo de Auto Scaling a que atenda aos critérios.

AWS CLI

Para desanexar e anexar uma instância, use os comandos de exemplo a seguir. Substitua cada *`user input placeholder`* por suas próprias informações.

Para desvincular uma instância de um grupo do Auto Scaling existente

1. Para descrever as instâncias atuais, use o describe-auto-scaling-instancescomando a seguir.

```
aws autoscaling describe-auto-scaling-instances \
  --query 'AutoScalingInstances[?AutoScalingGroupName==`my-asg`]'
```

O exemplo a seguir mostra a saída produzida quando você executa esse comando.

Anote o ID da instância que você pretende remover do grupo. Você precisa desse ID para a próxima etapa.

```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
```

```
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-0c20ac468fa3049e8",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-0787762faf1c28619",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-0f280a4c58d319a8a",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        }
    ]
}
```


2. Para desvincular uma instância e reduzir a capacidade desejada usando o seguinte comando detach-instances.

```
aws autoscaling detach-instances --instance-ids i-05b4f7d5be44822a6 \
  --auto-scaling-group-name my-asg
```

Para desanexar uma instância e diminuir a capacidade desejada, inclua a opção `--should-decrement-desired-capacity`.

```
aws autoscaling detach-instances --instance-ids i-05b4f7d5be44822a6 \
  --auto-scaling-group-name my-asg --should-decrement-desired-capacity
```

Agora você pode anexar a instância a um grupo do Auto Scaling diferente.

Como anexar uma instância a um grupo do Auto Scaling

1. Para anexar uma instância ao grupo do Auto Scaling use o seguinte comando attach-instances.

```
aws autoscaling attach-instances --instance-ids i-05b4f7d5be44822a6 --auto-
scaling-group-name my-asg-for-testing
```

2. Para verificar o tamanho do grupo Auto Scaling depois de anexar uma instância, use o comando a seguir. describe-auto-scaling-groups

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-names my-asg-
for-testing
```

O exemplo de resposta a seguir mostra que o grupo tem duas instâncias em execução, uma das quais é a instância que você anexou.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg-for-testing",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "2",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "MinSize": 1,
```

```
            "MaxSize": 5,
            "DesiredCapacity": 2,
            ...
            "Instances": [
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-05b4f7d5be44822a6",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService"
                },
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "2",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-00dcdfffdf5175890",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService"
                }
            ],
            ...
        }
    ]
}
```

Remover temporiamente instâncias do grupo do Auto Scaling

Você pode colocar uma instância que está no estado `InService` no estado `Standby`, atualize ou solucione problemas da instância e, em seguida, devolva a instância ao serviço. As instâncias que

estão em espera ainda fazem parte do grupo do Auto Scaling, mas não lidam ativamente com o tráfego do balanceador de carga.

Esse recurso ajuda a interromper e iniciar as instâncias ou reiniciá-las sem se preocupar com o término das instâncias do Amazon EC2 Auto Scaling como parte de suas verificações de saúde ou durante eventos de redução de escala na horizontal.

Por exemplo, você pode alterar a imagem de máquina da Amazon (AMI) para um grupo do Auto Scaling a qualquer momento alterando o modelo de execução ou a configuração de execução. Todas as instâncias subsequentes iniciadas pelo grupo do Auto Scaling usam essa AMI. No entanto, o grupo do Auto Scaling não atualiza as instâncias que estão em serviço atualmente. Você pode terminar essas instâncias e permitir que o Amazon EC2 Auto Scaling as substitua ou usar o recurso de atualização de instância para terminar e substituir as instâncias. Você também pode colocar as instâncias em espera, atualizar o software e, em seguida, colocar as instâncias de volta em serviço.

A desvinculação de instâncias de um grupo do Auto Scaling é semelhante a colocar instâncias em espera. Desanexar instâncias pode ser útil se você quiser anexá-las a um grupo diferente ou gerenciá-las como instâncias do EC2 independentes e, possivelmente, encerrá-las. Para obter mais informações, consulte Desvincular instâncias do seu grupo do Auto Scaling.

Conteúdo

- Como o estado de espera funciona
- Considerações
- Status de integridade de uma instância em um estado de espera
- Remover temporariamente uma instância definindo-a como em espera

Como o estado de espera funciona

O estado de espera funciona da seguinte forma para ajudar a remover temporariamente uma instância do seu grupo do Auto Scaling:

1. Você coloca uma instância no estado de espera. A instância permanece nesse estado até que você saia do estado de espera.
2. Se houver um grupo de destino de balanceador de carga ou um Classic Load Balancer anexado ao seu grupo do Auto Scaling, o registro da instância será cancelado no balanceador de carga. Se a descarga da conexão estiver habilitada para o balanceador de carga, o Elastic Load Balancing

aguardará 300 segundos por padrão antes de concluir o processo de cancelamento do registro, o que ajuda a solicitações em andamento a serem concluídas.

3. Você pode atualizar ou resolver problemas da instância.
4. Você devolve a instância para serviço saindo do estado de espera.
5. Se houver um grupo de destino de balanceador de carga ou um Classic Load Balancer anexado ao seu grupo do Auto Scaling, a instância será registrada no balanceador de carga.

Para obter mais informações sobre o ciclo de vida de instâncias em um grupo do Auto Scaling, consulte Ciclo de vida das instâncias do Amazon EC2 Auto Scaling.

Considerações

Veja a seguir algumas considerações ao mover instâncias para dentro e para fora do estado de espera:

- Ao colocar uma instância em espera, você pode diminuir a capacidade desejada por meio dessa operação ou mantê-la no mesmo valor.
  - Se você optar por não reduzir a capacidade desejada do grupo do Auto Scaling, o Amazon EC2 Auto Scaling iniciará uma instância para substituir a que está em espera. A intenção é ajudar você a manter a capacidade para a aplicação enquanto uma ou mais instâncias estão em espera.
  - Se você optar por diminuir a capacidade desejada do grupo do Auto Scaling, isso impedirá a execução de uma instância para substituir a que está em espera.
- Depois de colocar a instância novamente em serviço, a capacidade desejada é incrementada para refletir quantas instâncias estão no grupo do Auto Scaling.
- Para aumentar (e diminuir), a nova capacidade desejada deve estar entre o tamanho mínimo e máximo do grupo. Caso contrário, haverá falha na operação.
- Se, a qualquer momento, após colocar uma instância em espera ou retornar a instância ao serviço ao sair do estado de espera, descobrir que seu grupo do Auto Scaling não está balanceado entre as zonas de disponibilidade, o Amazon EC2 Auto Scaling compensará rebalanceando as zonas de disponibilidade, a menos que você suspenda o processo `AZRebalance`. Para obter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.
- Você é cobrado por instâncias que estão em estado de espera.

Status de integridade de uma instância em um estado de espera

O Amazon EC2 Auto Scaling não executa verificações de integridade em instâncias que estão em um estado de espera. Enquanto a instância está em estado de espera, seu status de integridade reflete o status que ela tinha antes de ser colocada em espera. O Amazon EC2 Auto Scaling não executa uma verificação de integridade na instância até você colocá-la em serviço.

Por exemplo, se você colocar uma instância íntegra em espera e, em seguida, terminá-la, o Amazon EC2 Auto Scaling continuará a relatar a instância como íntegra. Se você tentar colocar uma instância terminada que estava em espera em funcionamento novamente, o Amazon EC2 Auto Scaling executará uma verificação de integridade na instância, determinará que ela está sendo terminada e que não está íntegra e iniciará uma instância de substituição. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Remover temporariamente uma instância definindo-a como em espera

Use um dos procedimentos a seguir para tirar uma instância de serviço temporariamente, colocando-a em estado de espera.

Console

Para remover uma instância temporariamente

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), selecione uma instância.
4. Escolha Ações, Definir em espera.
5. Na caixa de diálogo Colocar em espera, mantenha a caixa de seleção Substituir instância para iniciar uma instância substituta. Desmarque a caixa de seleção para diminuir a capacidade desejada.
6. Quando a confirmação for solicitada, digite **`standby`** para confirmar a colocação da instância especificada no estado `Standby` e, em seguida, escolha Colocar em espera.
7. Você pode atualizar ou solucionar problemas de uma instância, conforme necessário. Quando tiver concluído, continue com a próxima etapa para retornar a instância para serviço.

8. Selecione a instância, escolha Ações, Definir como InService. Na caixa de InService diálogo Definir como, escolha Definir como InService.

AWS CLI

Para remover temporariamente uma instância de seu grupo do Auto Scaling, use os comandos de exemplo a seguir. Substitua cada *user input placeholder* por suas próprias informações.

Para remover uma instância temporariamente

1. Use o seguinte comando describe-auto-scaling-instances para identificar a instância a ser atualizada.

```
aws autoscaling describe-auto-scaling-instances \
  --query 'AutoScalingInstances[?AutoScalingGroupName==`my-asg`]'
```

O exemplo a seguir mostra a saída produzida quando você executa esse comando.

Anote o ID da instância que você pretende remover do grupo. Você precisa desse ID para a próxima etapa.

```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceId": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        ...
    ]
}
```

2. Mude a instância para o estado `Standby` usando o seguinte comando enter-standby. A opção `--should-decrement-desired-capacity` reduz a capacidade desejada para que o grupo do Auto Scaling não execute uma instância de substituição.

```
aws autoscaling enter-standby --instance-ids i-05b4f7d5be44822a6 \
  --auto-scaling-group-name my-asg --should-decrement-desired-capacity
```

O seguinte é um exemplo de resposta.

```
{
    "Activities": [
        {
            "ActivityId": "3b1839fe-24b0-40d9-80ae-bcd883c2be32",
            "AutoScalingGroupName": "my-asg",
            "Description": "Moving EC2 instance to Standby:
 i-05b4f7d5be44822a6",
            "Cause": "At 2023-12-15T21:31:26Z instance i-05b4f7d5be44822a6 was
 moved to standby
              in response to a user request, shrinking the capacity from 4 to
 3.",
            "StartTime": "2023-12-15T21:31:26.150Z",
            "StatusCode": "InProgress",
            "Progress": 50,
            "Details": "{\"Subnet ID\":\"subnet-c934b782\",\"Availability Zone
\":\"us-west-2a\"}"
        }
    ]
}
```

3. (Opcional) Verifique se a instância está em `Standby` usando o seguinte comando describe-auto-scaling-instances.

```
aws autoscaling describe-auto-scaling-instances --instance-
ids i-05b4f7d5be44822a6
```

O seguinte é um exemplo de resposta. Observe que o status da instância agora é `Standby`.

```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
```

```
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "Standby"
        },
        ...
    ]
}
```

4. Você pode atualizar ou solucionar problemas de uma instância, conforme necessário. Quando tiver concluído, continue com a próxima etapa para retornar a instância para serviço.
5. Coloque a instância de volta em serviço usando o seguinte comando exit-standby.

```
aws autoscaling exit-standby --instance-ids i-05b4f7d5be44822a6 --auto-scaling-
group-name my-asg
```

O seguinte é um exemplo de resposta.

```
{
    "Activities": [
        {
            "ActivityId": "db12b166-cdcc-4c54-8aac-08c5935f8389",
            "AutoScalingGroupName": "my-asg",
            "Description": "Moving EC2 instance out of Standby:
 i-05b4f7d5be44822a6",
            "Cause": "At 2023-12-15T21:46:14Z instance i-05b4f7d5be44822a6 was
 moved out of standby in
               response to a user request, increasing the capacity from 3 to
 4.",
            "StartTime": "2023-12-15T21:46:14.678Z",
            "StatusCode": "PreInService",
            "Progress": 30,
            "Details": "{\"Subnet ID\":\"subnet-c934b782\",\"Availability Zone
\":\"us-west-2a\"}"
        }
```

```
    ]
}
```

6. (Opcional) Verifique se a instância está de volta em serviço usando o seguinte comando `describe-auto-scaling-instances`.

```
aws autoscaling describe-auto-scaling-instances --instance-
ids i-05b4f7d5be44822a6
```

O seguinte é um exemplo de resposta. Observe que o status da instância é `InService`.

```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
       ...
    ]
}
```

Excluir infraestrutura do Auto Scaling

Para excluir completamente sua infraestrutura de escalabilidade, execute as tarefas a seguir.

Tarefas

- Excluir seu grupo do Auto Scaling
- (Opcional) Excluir a configuração de execução
- (Opcional) Excluir o modelo de execução

- (Opcional) Excluir o balanceador de carga e grupos de destino
- (Opcional) Excluir CloudWatch alarmes
- Configure a proteção contra exclusão para seus recursos do Amazon EC2 Auto Scaling

Excluir seu grupo do Auto Scaling

Quando você exclui um grupo do Auto Scaling, seus valores desejado, mínimo e máximo são definidos como 0. Como resultado, as instâncias são encerradas. A exclusão de uma instância também exclui os logs ou os dados associados e todos os volumes na instância. Se não quiser terminar uma ou mais instâncias, poderá desvinculá-las antes de excluir o grupo do Auto Scaling. Se o grupo tiver políticas de escalabilidade, a exclusão do grupo excluirá as políticas, as ações de alarme subjacentes e qualquer alarme que não tenha mais uma ação associada.

Para excluir seu grupo do Auto Scaling (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling e escolha Ações, Excluir.
3. Quando a confirmação for solicitada, digite **`delete`** para confirmar a exclusão do grupo do Auto Scaling especificado e, em seguida, escolha Excluir.

   Um ícone de carregamento na coluna Name (Nome) indica que o grupo do Auto Scaling está sendo excluído. As colunas Desejado, Mínimo e Máximo mostram `0` instâncias para o grupo do Auto Scaling. São necessários alguns minutos para encerrar a instância e excluir o grupo. Atualize a lista para ver o estado atual.

Excluir seu grupo do Auto Scaling (AWS CLI)

Use o delete-auto-scaling-groupcomando a seguir para excluir o grupo Auto Scaling. Essa operação não funciona se o grupo tiver alguma instância do EC2; é somente para grupos com zero instâncias.

```
aws autoscaling delete-auto-scaling-group --auto-scaling-group-name my-asg
```

Se o grupo tiver instâncias ou ações de escalabilidade em andamento, use o comando delete-auto-scaling-group com a opção `--force-delete`. Isso também encerrará as instâncias do EC2. Quando você exclui um grupo do Auto Scaling do console do Amazon EC2 Auto Scaling, o console usa essa operação para encerrar qualquer instância do EC2 e excluir o grupo ao mesmo tempo.

```
aws autoscaling delete-auto-scaling-group --auto-scaling-group-name my-asg --force-
delete
```

(Opcional) Excluir a configuração de execução

Você pode ignorar esta etapa para manter a configuração de execução para uso futuro.

Para excluir a configuração de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, em Ajuste de escala automático, escolha Grupos do Auto Scaling.
3. Escolha Configurações de execução na parte superior da página. Quando solicitada a confirmação, escolha Exibir configurações de execução para confirmar que você deseja exibir a página de Configurações de execução.
4. Selecione sua configuração de inicialização e escolha Ações, Excluir configuração de inicialização.
5. Quando a confirmação for solicitada, escolha Excluir.

Para excluir a configuração de ativação (AWS CLI)

Use o seguinte comando delete-launch-configuration:

```
aws autoscaling delete-launch-configuration --launch-configuration-name my-launch-
config
```

(Opcional) Excluir o modelo de execução

Você pode excluir o modelo de execução ou apenas uma versão do seu modelo de execução. Ao excluir um modelo de execução, todas as suas versões são excluídas.

É possível ignorar esta etapa para manter o modelo de execução para uso futuro.

Para excluir seu modelo de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, escolha Instances e, em seguida, Launch Templates.

3. Selecione o modelo de execução e, depois, execute uma das seguintes ações:

   - Escolha Actions (Ações), Delete template (Excluir modelo). Quando a confirmação for solicitada, digite **Delete** para confirmar a exclusão do modelo de execução especificado e, em seguida, escolha Excluir.
   - Escolha Actions (Ações), Delete template version (Excluir versão do modelo). Selecione a versão a ser excluída e escolha Delete (Excluir).

Para excluir o modelo de execução (AWS CLI)

Use o comando delete-launch-template a seguir para excluir seu modelo e todas as suas versões.

```
aws ec2 delete-launch-template --launch-template-id lt-068f72b72934aff71
```

Como alternativa, você pode usar o comando delete-launch-template-versions para excluir uma versão específica de um modelo de execução.

```
aws ec2 delete-launch-template-versions --launch-template-id lt-068f72b72934aff71 --
versions 1
```

(Opcional) Excluir o balanceador de carga e grupos de destino

Ignore esta etapa se seu grupo do Auto Scaling não estiver associado a um balanceador de carga Elastic Load Balancing ou se desejar manter o balanceador de carga para uso futuro.

Para excluir o balanceador de carga (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, em Load Balancing (Balanceamento de carga), escolha Load balancers (Balanceadores de carga).
3. Selecione o balanceador de carga e Actions (Ações), Delete (Excluir).
4. Quando a confirmação for solicitada, escolha Sim, excluir.

Para excluir o grupo de destino (console)

1. No painel de navegação, em Load Balancing (Balanceamento de carga), escolha Grupos de destino.

2. Selecione o grupo de destino e escolha Actions (Ações), Delete (Excluir).
3. Quando a confirmação for solicitada, escolha Sim, excluir.

Para excluir o balanceador de carga associado ao grupo do Auto Scaling (AWS CLI)

Para balanceadores de carga de aplicativos e balanceadores de carga de rede, use os comandos a seguir delete-load-balancer. delete-target-group

```
aws elbv2 delete-load-balancer --load-balancer-arn my-load-balancer-arn
aws elbv2 delete-target-group --target-group-arn my-target-group-arn
```

Para balanceadores de carga clássicos, use o delete-load-balancercomando a seguir.

```
aws elb delete-load-balancer --load-balancer-name my-load-balancer
```

(Opcional) Excluir CloudWatch alarmes

Para excluir os CloudWatch alarmes associados ao seu grupo de Auto Scaling, conclua as etapas a seguir. Por exemplo, você pode ter alarmes associados à escalabilidade por etapas ou às políticas de escalabilidade simples.

> **Note**
>
> A exclusão de um grupo do Auto Scaling exclui automaticamente CloudWatch os alarmes que o Amazon EC2 Auto Scaling gerencia para uma política de escalabilidade de rastreamento alvo.

Você pode pular essa etapa se o grupo do Auto Scaling não estiver associado a CloudWatch nenhum alarme ou se quiser manter os alarmes para uso futuro.

Para excluir os CloudWatch alarmes (console)

1. Abra o CloudWatch console em https://console.aws.amazon.com/cloudwatch/.
2. No painel de navegação, escolha Alarms (Alarmes).
3. Selecione os alarmes e escolha Action (Ação), Delete (Excluir).
4. Quando a confirmação for solicitada, escolha Excluir.

Para excluir os CloudWatch alarmes ()AWS CLI

Use o comando delete-alarms. É possível excluir um ou mais alarmes por vez. Por exemplo, use o comando a seguir para excluir os alarmes `Step-Scaling-AlarmHigh-AddCapacity` e `Step-Scaling-AlarmLow-RemoveCapacity`.

```
aws cloudwatch delete-alarms --alarm-name Step-Scaling-AlarmHigh-AddCapacity Step-Scaling-AlarmLow-RemoveCapacity
```

Configure a proteção contra exclusão para seus recursos do Amazon EC2 Auto Scaling

Proteja sua infraestrutura do Amazon EC2 Auto Scaling contra exclusão acidental configurando várias camadas de proteção. O Auto Scaling fornece várias abordagens para evitar a exclusão indesejada de recursos para seus grupos de Auto Scaling e para as instâncias do Amazon EC2 que ele gerencia.

Conteúdo

- Configurar a proteção contra exclusão de grupos do Auto Scaling
- Controle as permissões de exclusão com políticas do IAM

Configurar a proteção contra exclusão de grupos do Auto Scaling

A proteção contra exclusão é uma configuração em nível de recurso que impede que seu grupo Amazon EC2 Auto Scaling seja excluído acidentalmente. Quando ativada, a proteção contra exclusão impede que a operação da DeleteAutoScalingGroupAPI seja bem-sucedida, exigindo que você primeiro atualize a configuração de proteção contra exclusão para um nível menos restritivo antes de excluir o grupo Auto Scaling.

O Amazon EC2 Auto Scaling oferece três níveis de proteção contra exclusão:

Nenhum (padrão)

> Nenhuma proteção de exclusão está ativada, o que significa que seu grupo do Auto Scaling pode ser excluído com ou sem o uso `ForceDelete` da opção. Quando `ForceDelete` usadas, todas as instâncias do Amazon EC2 gerenciadas pelo seu grupo do Auto Scaling também serão encerradas à força sem executar ganchos do ciclo de vida da rescisão.

Impedir a exclusão forçada

Seu grupo de Auto Scaling não pode ser excluído ao usar a `ForceDelete` opção. Essa configuração permite a exclusão de grupos vazios do Auto Scaling (grupos sem instâncias). Essa opção é recomendada para cargas de trabalho de produção nas quais você deseja evitar o encerramento em massa de instâncias, mas permitir a limpeza de grupos vazios.

Evitar todas as exclusões

Seu grupo de Auto Scaling não pode ser excluído, independentemente de a `ForceDelete` opção ser usada. Essa opção oferece a proteção mais forte contra exclusão acidental. É necessário desativar explicitamente a proteção contra exclusão antes que seu grupo do Auto Scaling possa ser excluído. Isso é recomendado para grupos de missão crítica do Auto Scaling que raramente ou nunca devem ser excluídos.

Como funciona a proteção contra exclusão

Quando você tenta a operação da DeleteAutoScalingGroupAPI com a proteção de exclusão ativada:

1. O Amazon EC2 Auto Scaling valida a configuração de proteção contra exclusão antes de processar a solicitação.
2. Se o nível de proteção de exclusão configurado bloquear a tentativa de exclusão, o Amazon EC2 Auto Scaling retornará um. `ValidationError`
3. Seu grupo de Auto Scaling e suas instâncias do Amazon EC2 permanecem inalterados.
4. Você deve atualizar a configuração de proteção contra exclusão para um nível menos restritivo antes de excluir seu grupo de Auto Scaling.

A proteção contra exclusão não impede outras operações, como:

- Atualização da configuração do grupo Auto Scaling.
- Encerramento de instâncias individuais.
- Operações de escalonamento (manuais ou automáticas).
- Suspender ou retomar processos.

Para obter mais informações sobre como lidar normalmente com o encerramento de instâncias, consulte. Projete suas aplicações para lidar com a terminação de instâncias sem problemas

Configurar a proteção contra exclusão

Você pode definir a proteção contra exclusão ao criar um grupo de Auto Scaling ou atualizar a configuração em um grupo de Auto Scaling existente.

Console

Para criar um grupo de Auto Scaling com proteção contra exclusão

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Selecione Criar grupo do Auto Scaling.
3. Conclua as etapas de configuração do seu grupo de Auto Scaling.
4. Na página Configurar tamanho e escala do grupo, expanda Configurações adicionais.
5. Para proteção contra exclusão de grupos do Auto Scaling, escolha o nível de proteção desejado:
   - Nenhuma - Sem proteção contra exclusão (padrão)
   - Evitar exclusão forçada - Bloquear operações de exclusão forçada
   - Evitar todas as exclusões - Bloquear todas as operações de exclusão
6. Conclua as etapas restantes para criar seu grupo de Auto Scaling.

Para atualizar a proteção contra exclusão em um grupo existente do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.
3. Selecione Ações, Editar.
4. Em Configurações adicionais, atualize a configuração de proteção contra exclusão de grupos do Auto Scaling.
5. Selecione Atualizar.

AWS CLI

Para criar um grupo de Auto Scaling com proteção contra exclusão

Use o comando create-auto-scaling-group com o parâmetro `--deletion-protection`:

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateName=my-template,Version='$Latest' \
    --min-size 1 \
    --max-size 5 \
    --desired-capacity 2 \
    --vpc-zone-identifier "subnet-12345678,subnet-87654321" \
    --deletion-protection prevent-force-deletion
```

Os valores válidos para `--deletion-protection` são: none | `prevent-force-deletion` | `prevent-all-deletion`

Para atualizar a proteção contra exclusão em um grupo existente do Auto Scaling

Use o comando update-auto-scaling-group:

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --deletion-protection prevent-all-deletion
```

Para desativar a proteção contra exclusão

Defina a proteção contra exclusão comonone:

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --deletion-protection none
```

Para verificar o status da proteção contra exclusão

Use o comando describe-auto-scaling-groups:

```
aws autoscaling describe-auto-scaling-groups \
    --auto-scaling-group-names my-asg
```

Controle as permissões de exclusão com políticas do IAM

Use políticas AWS Identity and Access Management (IAM) para controlar quais usuários e funções podem excluir grupos do Auto Scaling. Os controles baseados em IAM fornecem uma camada adicional de segurança ao restringir as permissões no nível da identidade.

As políticas do IAM são particularmente úteis quando você deseja:

- Permita que diferentes usuários tenham diferentes níveis de acesso às operações do Auto Scaling.
- Impeça que usuários específicos usem a `ForceDelete` opção, mesmo que possam realizar outras operações de Auto Scaling.
- Restrinja as permissões de exclusão a grupos específicos do Auto Scaling.

A política a seguir permite a exclusão de um grupo do Auto Scaling somente se o grupo tiver a tag **`environment=development`**.

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": "autoscaling:DeleteAutoScalingGroup",
      "Resource": "*",
      "Condition": {
          "StringEquals": { "aws:ResourceTag/environment": "development" }
      }
   }]
}
```

A política a seguir usa a chave de `autoscaling:ForceDelete` condição para controlar o acesso à ação `DeleteAutoScalingGroup` da API. Isso pode impedir que determinados usuários usem a `ForceDelete` operação, que encerra todas as instâncias do Amazon EC2 dentro de um grupo de Auto Scaling.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [{
        "Effect": "Deny",
        "Action": "autoscaling:DeleteAutoScalingGroup",
        "Resource": "*",
        "Condition": {
```

```
            "Bool": {
                "autoscaling:ForceDelete": "true"
            }
        }
    }]
}
```

Como alternativa, se você não estiver usando chaves de condição para controlar o acesso aos grupos do Auto Scaling, você pode especificar o ARNs número de recursos no `Resource` elemento para controlar o acesso.

A política a seguir dá aos usuários permissões para usar a ação da `DeleteAutoScalingGroup` API, mas somente para grupos de Auto Scaling cujo nome começa com. **`devteam-`**

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "autoscaling:DeleteAutoScalingGroup",
            "Resource": "arn:aws:autoscaling:us-
east-1:111122223333:autoScalingGroup:*:autoScalingGroupName/devteam-*"
        }
    ]
}
```

Você também pode especificar vários ARNs colocando-os em uma lista. A inclusão da UUID garante que o acesso seja concedido ao grupo do Auto Scaling específico. O UUID de um novo grupo é diferente do UUID de um grupo excluído com o mesmo nome.

```
"Resource": [
    "arn:aws:autoscaling:region:account-
id:autoScalingGroup:uuid:autoScalingGroupName/devteam-1",
    "arn:aws:autoscaling:region:account-
id:autoScalingGroup:uuid:autoScalingGroupName/devteam-2",
    "arn:aws:autoscaling:region:account-
id:autoScalingGroup:uuid:autoScalingGroupName/devteam-3"
```

```
]
```

Para obter exemplos adicionais de políticas do IAM para o Amazon EC2 Auto Scaling, incluindo políticas que controlam as permissões de exclusão, consulte. Exemplos de políticas baseadas em identidade

Recicle as instâncias em seu grupo do Auto Scaling

O Amazon EC2 Auto Scaling oferece recursos que permitem substituir as instâncias do Amazon EC2 em seu grupo do Auto Scaling após fazer atualizações, como novos tipos de instância ou um modelo de execução atualizado. Também ajuda a simplificar as atualizações oferecendo a opção de incluí-las na mesma operação que substitui as instâncias.

Esta seção inclui informações para ajudar você a fazer o seguinte:

- Iniciar uma atualização de instância para substituir instâncias no grupo do Auto Scaling.
- Declarar atualizações específicas que descrevem uma configuração desejada e atualizar o grupo do Auto Scaling para a configuração desejada.
- Pular a substituição de instâncias já atualizadas.
- Usar pontos de verificação para substituir instâncias em fases e realizar verificações em suas instâncias em pontos específicos.
- Use o tempo de incorporação para fazer uma pausa no final de uma atualização da instância para validar a integridade da instância.
- Receber notificações por e-mail quando um ponto de verificação for atingido.
- Utilize uma reversão para restaurar o grupo do Auto Scaling para a configuração que ele estava usando anteriormente.
- Reverta automaticamente se a atualização da instância falhar por algum motivo ou se algum CloudWatch alarme da Amazon que você especificar entrar no `ALARM` estado.
- Limitar a vida útil das instâncias para fornecer versões de software consistentes e configurações de instância em todo o grupo do Auto Scaling.
- Substitua os volumes raiz sem encerrar instâncias, mantendo as interfaces de rede, os volumes não raiz e as políticas do IAM.

Conteúdo

- Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling
- Substituir instâncias do Auto Scaling com base na vida útil máxima da instância

Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling

Você inicia as atualizações de instâncias para atualizar as instâncias em seu grupo do Auto Scaling. Esse recurso pode ser útil quando uma alteração na configuração exige que você substitua instâncias ou seus volumes raiz, especialmente se seu grupo de Auto Scaling contiver um grande número de instâncias.

Algumas situações em que uma atualização de instância pode ajudar incluem:

- Implantação de uma nova Imagem de Máquina da Amazon (AMI) ou script de dados do usuário em seu grupo do Auto Scaling. Você pode criar um novo modelo de execução com as alterações e, em seguida, usar uma atualização de instância para implementar as atualizações imediatamente.
- Migrar suas instâncias para novos tipos de instância para aproveitar as melhorias e otimizações mais recentes.
- Mudança dos grupos do Auto Scaling do uso de uma configuração de execução para o uso de um modelo de execução. Você pode copiar suas configurações de execução para modelos de execução e, em seguida, usar uma atualização de instância para atualizar suas instâncias para os novos modelos. Para obter mais informações sobre a migração para modelos de lançamento, consulte Migre seus grupos do Auto Scaling para modelos de execução.
- Aplique patches de segurança ou atualizações de software, preservando o estado da instância de longa duração e evitando restrições de capacidade com tipos de instância especializados, como instâncias de GPU ou Mac.

Conteúdo

- Como funciona a atualização de uma instância em um grupo do Auto Scaling
- Entender os valores padrão de uma atualização de instância
- Inicie uma atualização de instância usando o ou Console de gerenciamento da AWS AWS CLI
- Monitore a atualização de uma instância usando o ou Console de gerenciamento da AWS AWS CLI
- Substitua os volumes raiz durante a atualização da instância
- Cancelar uma atualização de instância usando o ou Console de gerenciamento da AWS AWS CLI
- Desfazer alterações com uma reversão manual ou automática
- Usar uma atualização de instância com opção de ignorar correspondência

- Adicionar pontos de verificação a uma atualização de instância

Como funciona a atualização de uma instância em um grupo do Auto Scaling

Este tópico descreve como funciona uma atualização de instância e apresenta os principais conceitos que você precisa entender para usá-la de forma eficaz.

Conteúdo

- Como funcionam
- Conceitos principais
- Período de carência da verificação de integridade
- Compatibilidade de tipo de instância
- Limitações

Como funcionam

Para atualizar instâncias em um grupo do Auto Scaling, você pode definir uma nova configuração que contenha a versão mais recente da sua aplicação e quaisquer outras atualizações que queira fazer.

A atualização de instâncias oferece suporte a duas estratégias para atualizar instâncias:

- Estratégia contínua (padrão) — encerra instâncias e inicia novas em lotes de acordo com suas preferências. Isso garante que seu grupo de Auto Scaling mantenha a capacidade e a disponibilidade desejadas durante todo o processo de atualização.
- Substitua a estratégia do volume raiz — atualiza as instâncias substituindo somente o volume raiz sem encerrar a instância. Isso preserva as interfaces de rede da instância, os volumes não raiz do EBS e os dados do armazenamento da instância.

Requisitos para a estratégia de substituição do volume raiz:

- Seu grupo de Auto Scaling deve usar uma política de instâncias mistas
- Todas as substituições na política de instâncias mistas devem especificar um `ImageId`
- AMIs deve conter somente um único volume raiz

- Todas as instâncias devem corresponder à configuração do modelo de lançamento do grupo
- Você deve iniciar a atualização da instância com a configuração desejada, com uma política de instâncias mistas contendo `ImageId` substituições.

Quando o skip matching está ativado, o Auto Scaling compara o ID da AMI atual de cada instância com a IDs AMI na configuração desejada. Ele substitui apenas as instâncias em que a AMI IDs não coincide, permitindo que você ignore instâncias que já estão atualizadas.

Execute uma atualização de instância

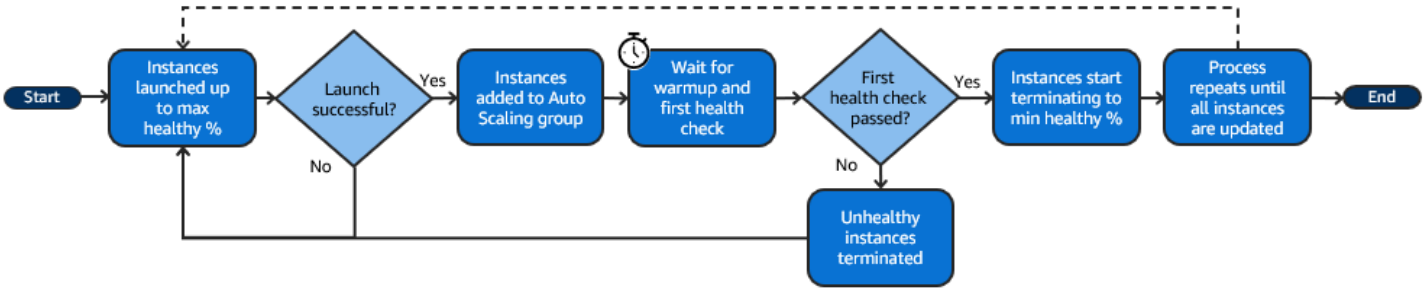
Inicie uma atualização da instância para substituir as existentes por novas com base nessa configuração.

1. Crie um novo modelo de execução ou atualize o modelo existente com as alterações de configuração desejadas, como uma nova imagem de máquina da Amazon (AMI). Para obter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.
2. Inicie a atualização da instância usando o console AWS CLI ou SDK do Amazon EC2 Auto Scaling:
   - Especifique o novo modelo de execução ou a versão do modelo de execução que você criou. Isso será usado para iniciar novas instâncias.
   - Defina as porcentagens mínima e máxima de integridade. Isso controla quantas instâncias são substituídas simultaneamente e se as novas instâncias são iniciadas antes de terminar as antigas.
   - Configure todas as definições opcionais, como:
     - Pontos de controle: Pause a atualização da instância após uma determinada porcentagem de substituições para verificar o progresso.
     - Tempo de incorporação: faça uma pausa no final da atualização da instância para validar a integridade da instância antes que a atualização da instância seja considerada concluída.
     - Ignorar correspondência: Compare as instâncias antigas com a nova configuração e substitua apenas as que não corresponderem. Quando você inicia uma atualização de instância no console, a ignorância de correspondência está ativada por padrão.
     - Vários tipos de instância: aplique uma política de instâncias mistas nova ou atualizada como parte da configuração desejada.

Quando a atualização da instância tiver sido iniciada, o Amazon EC2 Auto Scaling começará a fazer as ações a seguir:

- Substitua as instâncias em lotes com base nas porcentagens mínimas e máximas de integridade.
- Execute as novas instâncias primeiro, antes de terminar as antigas, se a porcentagem mínima de integridade estiver definida como 100%. Isso garante que sua capacidade desejada seja mantida o tempo todo.
- Verifique o estado de integridade das instâncias e dê a elas tempo para se aquecerem antes de substituir mais instâncias.
- Termine e substitua as instâncias consideradas não íntegras.
- Atualize automaticamente as definições do grupo do Auto Scaling com as novas alterações de configuração depois que a atualização da instância for bem-sucedida.
- Substitua `InService` instâncias antes das instâncias que estão em um pool ativo.

O fluxograma a seguir ilustra o comportamento de execução antes de terminação quando você define a porcentagem mínima de integridade como 100%.



> **Note**
>
> Especifique apenas a porcentagem máxima íntegra para uma atualização de instância se você não tiver definido uma política de manutenção de instância ou se precisar substituí-la. Para obter mais informações, consulte Políticas de manutenção de instância.
> Da mesma forma, você só precisa especificar o período de aquecimento da instância para fazer a atualização se não tiver ativado o aquecimento padrão ou se precisar substituir o padrão. Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Conceitos principais

Antes de começar, familiarize-se com os seguintes conceitos básicos de atualização de instância:

Percentual mínimo de integridade

A porcentagem mínima de integridade é a porcentagem da capacidade desejada para se manter em serviço, íntegra e pronta para uso durante uma atualização de instância para que a atualização possa continuar. Por exemplo, se a porcentagem mínima de integridade for 90% e a porcentagem máxima de integridade for 100%, 10% da capacidade será substituída por vez. Se as novas instâncias não passarem nas verificações de integridade, o Amazon EC2 Auto Scaling as encerrará e substituirá. Se a atualização da instância não puder iniciar nenhuma instância íntegra, ela eventualmente falhará, deixando os outros 90% do grupo intactos. Se as novas instâncias permanecerem íntegras e terminarem o período de aquecimento, o Amazon EC2 Auto Scaling poderá continuar a substituir outras instâncias.

A atualização de instância pode substituir uma instância por vez, várias por vez ou todas de uma vez. Para substituir uma instância por vez, defina a porcentagem mínima e máxima de integridade como 100%. Isso altera o comportamento de uma atualização de instância para ser iniciada antes do encerramento, o que evita que a capacidade do grupo fique abaixo de 100% da capacidade desejada. Para substituir todas as instâncias de uma vez, defina uma porcentagem mínima de integridade de 0%.

Porcentagem máxima de integridade

A porcentagem máxima íntegra é a porcentagem da capacidade desejada que seu grupo do Auto Scaling pode aumentar ao substituir instâncias. A diferença entre o mínimo e o máximo não pode ser maior que 100. Um intervalo maior aumenta o número de instâncias que podem ser substituídas ao mesmo tempo.

Aquecimento da instância

O aquecimento da instância é o período de tempo desde a mudança do estado de uma nova instância até o momento em que a `InService` inicialização é considerada concluída. Durante uma atualização de instância, se as instâncias passam na verificação de integridade, o Amazon EC2 Auto Scaling não avança imediatamente para substituir a próxima instância após determinar que uma instância recém-iniciada está íntegra. Ele aguarda o período de aquecimento antes de começar a substituir a próxima instância. Isso pode ser útil quando a aplicação ainda precisar de um tempo de inicialização antes de responder às solicitações.

O aquecimento da instância funciona da mesma forma que o aquecimento de instâncias padrão. Portanto, as mesmas considerações de escalabilidade são aplicadas. Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Configuração desejada

A configuração desejada é a nova configuração que você deseja que o Amazon EC2 Auto Scaling implante no grupo do Auto Scaling. Por exemplo, você pode especificar um novo modelo de execução e novos tipos de instância para suas instâncias. Durante uma atualização de instância, o Amazon EC2 Auto Scaling atualiza o grupo do Auto Scaling para a configuração desejada. Se um evento aumento da escala na horizontal ocorrer durante uma atualização de instância, o Amazon EC2 Auto Scaling iniciará novas instâncias com a configuração desejada em vez das configurações atuais do grupo. Depois que a atualização de instância tem êxito, o Amazon EC2 Auto Scaling atualiza as configurações do grupo do Auto Scaling para refletir a nova configuração desejada que você especificou como parte da atualização de instância.

Ignorar correspondência

Ignorar a correspondência diz ao Amazon EC2 Auto Scaling para ignorar as instâncias que já tenham as atualizações mais recentes. Assim, você não substituirá mais instâncias do que o necessário. Isso é útil quando você deseja garantir que o grupo do Auto Scaling usará uma versão específica de seu modelo de execução e substituirá apenas as instâncias que usam outra versão.

Pontos de verificação

Um ponto de verificação é um ponto no tempo em que a atualização de instância é interrompida por um período especificado. Uma atualização de instância pode conter vários pontos de verificação. O Amazon EC2 Auto Scaling emite eventos para cada ponto de verificação. Portanto, você pode adicionar uma EventBridge regra para enviar os eventos para um destino, como o Amazon SNS, para ser notificado quando um ponto de verificação for alcançado. Depois que um ponto de verificação é atingido, você tem a oportunidade de verificar sua implantação. Se algum problema for identificado, você poderá cancelar a atualização de instância ou revertê-la. A capacidade de implantar atualizações em fases é um benefício fundamental dos pontos de verificação. Se você não usar pontos de verificação, as substituições continuas serão executadas ininterruptamente.

Para saber mais sobre todas as configurações padrão que você pode definir ao iniciar uma atualização de instância, consulte Entender os valores padrão de uma atualização de instância.

Período de carência da verificação de integridade

O Amazon EC2 Auto Scaling determina se a instância está íntegra com base no status das verificações de integridade que o grupo do Auto Scaling usa. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Para garantir que essas verificações de integridade comecem o mais rápido possível, não defina um período de carência da verificação de integridade do grupo muito alto, mas alto o suficiente para que suas verificações de integridade do Elastic Load Balancing consigam determinar se um destino está disponível para lidar com solicitações. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

Compatibilidade de tipo de instância

Antes de alterar o tipo de instância, convém verificar se ela funciona com seu modelo de execução. Isso confirma a compatibilidade com a AMI especificada. Por exemplo, digamos que você iniciou suas instâncias originais com base em uma AMI paravirtual (PV), mas deseja alterar para um tipo de instância da geração atual que tenha suporte apenas em uma AMI de máquina virtual (HVM). Nesse caso, é necessário usar uma AMI HVM no modelo de execução.

Para confirmar a compatibilidade do tipo de instância sem iniciar instâncias, use o comando run-instances com a opção `--dry-run`, conforme mostrado no exemplo a seguir.

```
aws ec2 run-instances --launch-template LaunchTemplateName=my-template,Version='1' --
dry-run
```

Para obter informações sobre como a compatibilidade é determinada, consulte Compatibilidade para alterar o tipo de instância no Guia do usuário do Amazon EC2 para instâncias do Linux.

Limitações

- Duração total: o tempo máximo que uma atualização de instância pode permanecer ativamente substituindo instâncias é de 14 dias.
- Diferença no comportamento específico de grupos ponderados: se um grupo de instâncias mistas estiver configurado com um peso de instância maior ou igual à capacidade desejada do grupo, o Amazon EC2 Auto Scaling poderá substituir todas as instâncias `InService` de uma só vez. Para evitar essa situação, siga a recomendação do tópico Configurar um grupo do Auto Scaling para usar ponderação da instância. Especifique uma capacidade desejada que seja maior do que seu maior peso ao usar pesos com seu grupo do Auto Scaling.

- Tempo limite de uma hora: quando uma atualização de instância é incapaz de continuar fazendo substituições porque a aplicação está aguardando para substituir instâncias em espera ou protegidas contra a redução da escala na horizontal, ou se as novas instâncias não passarem nas verificações de integridade, o Amazon EC2 Auto Scaling continuará fazendo novas tentativas por uma hora. Ele também fornece uma mensagem de status para ajudar você a resolver o problema. Se o problema persistir após uma hora, a operação falhou. A intenção é garantir tempo para a recuperação em caso de um problema temporário.
- Implementação de código por meio de dados do usuário: ignorar a correspondência não verifica as alterações de código implementadas a partir de um script de dados do usuário. Se você usar os dados do usuário para extrair novos códigos e instalar essas atualizações em novas instâncias, recomendamos que desative a ignorância de correspondência para garantir que todas as instâncias recebam o código mais recente, mesmo sem uma atualização da versão do modelo de execução.
- Restrição de atualização: se você tentar atualizar o modelo de execução, a configuração de execução ou a política de instâncias mistas de um grupo do Auto Scaling enquanto uma atualização de instância com uma configuração desejada estiver ativa, a solicitação falhará com o erro de validação a seguir: `An active instance refresh with a desired configuration exists. All configuration options derived from the desired configuration are not available for update while the instance refresh is active.`
- Seleção de instância com base em atributos: caso seu grupo do Auto Scaling use seleção de instância baseada em atributos (especificando `InstanceRequirements` em uma política de instâncias mistas), não haverá suporte aos parâmetros de atualização de instância a seguir:
  - SkipMatching- A atualização da instância falhará se você iniciá-la com o `SkipMatching` parâmetro em um grupo de Auto Scaling que tenha `InstanceRequirements` sido configurado.
  - DesiredConfiguration- A atualização da instância falhará se você iniciá-la com o `DesiredConfiguration` parâmetro em um grupo de Auto Scaling que tenha `InstanceRequirements` sido configurado.

  Se você precisar realizar uma atualização de instância em um grupo do Auto Scaling com seleção de instância baseada em atributos, inicie a atualização da instância sem esses parâmetros.
- A substituição do volume raiz não é compatível com o Elastic Load Balancing.

Entender os valores padrão de uma atualização de instância

Antes de iniciar uma atualização de instância, é possível personalizar diversas preferências que afetam a atualização de instância. Alguns padrões de preferência são diferentes dependendo se você usa o console ou a linha de comando (AWS CLI ou AWS SDK).

A tabela a seguir lista os valores padrão das configurações de atualização de instância.

Configuração	AWS CLI ou AWS SDK	Console do Amazon EC2 Auto Scaling
CloudWatch alarme	Desativado (nulo)	Desabilitado
Reversão automática	Desabilitado (`false`)	Desabilitado
Hora de assar	Zero	Zero
Pontos de verificação	Desabilitado (`false`)	Desabilitado
Atraso no ponto de verificação	1 hora (`3600` segundos)	1 hora
Aquecimento da instância	O aquecimento de instâncias padrão, se estiver definido, ou o período de carência da verificação de integridade, se não estiver.	O aquecimento de instâncias padrão, se estiver definido, ou o período de carência da verificação de integridade, se não estiver.
Porcentagem máxima de integridade	Varia com base em sua política de manutenção de instâncias. Se não houver política de manutenção de instâncias, o padrão é 100% (nulo).	Varia com base em sua política de manutenção de instâncias. Se não houver política de manutenção de instâncias, o padrão é 100% (nulo).
Percentual mínimo de integridade	Varia com base em sua política de manutenção de instâncias. Se não houver política de manutenção de instâncias, o padrão é 90%.	Varia com base em sua política de manutenção de instâncias. Se não houver política de manutenção de instâncias, o padrão é 90%.

Configuração	AWS CLI ou AWS SDK	Console do Amazon EC2 Auto Scaling
Instâncias protegidas contra redução da escala na horizontal	Aguardar	Ignorar
Ignorar correspondência	Desabilitado (`false`)	Habilitado
Instâncias em espera	Aguardar	Ignorar

Segue uma descrição de cada configuração:

CloudWatch alarme (**`AlarmSpecification`**)

A especificação do CloudWatch alarme. CloudWatch os alarmes podem ser usados para identificar quaisquer problemas e falhar na operação se um alarme entrar no `ALARM` estado. Para obter mais informações, consulte Iniciar uma atualização de instância com reversão automática.

Reversão automática (**`AutoRollback`**)

Controla se o Amazon EC2 Auto Scaling reverte o grupo do Auto Scaling para sua configuração anterior se a atualização da instância falhar. Para obter mais informações, consulte Desfazer alterações com uma reversão manual ou automática.

Tempo de incorporação (**`BakeTime`**)

A quantidade de tempo de espera no final de uma atualização da instância antes que a atualização da instância seja considerada concluída.

Pontos de verificação (**`CheckpointPercentages`**)

Controla se o Amazon EC2 Auto Scaling substitui instâncias em fases. Isso é útil se você precisar realizar verificações em suas instâncias antes de substituir todas as instâncias. Para obter mais informações, consulte Adicionar pontos de verificação a uma atualização de instância.

Atraso no ponto de verificação (**`CheckpointDelay`**)

A quantidade de tempo, em segundos, para aguardar após um ponto de verificação antes de continuar. Para obter mais informações, consulte Adicionar pontos de verificação a uma atualização de instância.

Aquecimento da instância (**`InstanceWarmup`**)

Um período, em segundos, durante o qual o Amazon EC2 Auto Scaling espera até que uma nova instância seja considerada como inicialização concluída antes de substituir a próxima instância. Se você já definiu corretamente um aquecimento de instâncias padrão para o grupo do Auto Scaling, não é necessário alterar o aquecimento da instância (a menos que deseje substituir o padrão). Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Porcentagem máxima de integridade (**`MaxHealthyPercentage`**)

A porcentagem da capacidade desejada do grupo do Auto Scaling que seu grupo pode aumentar ao substituir instâncias.

Percentual mínimo de integridade (**`MinHealthyPercentage`**)

A porcentagem da capacidade desejada do grupo do Auto Scaling que deve estar em serviço, íntegra e pronta para uso antes que a operação possa continuar.

Instâncias protegidas contra redução da escala na horizontal (**`ScaleInProtectedInstances`**)

Controla o que o Amazon EC2 Auto Scaling faz se forem encontradas instâncias protegidas contra redução de escala. Para obter mais informações sobre essas instâncias, consulte Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância.

O Amazon EC2 Auto Scaling fornece estas opções:

- Substituir (`Refresh`): substitui instâncias que estão protegidas contra a redução da escala horizontalmente.
- Ignorar (`Ignore`): ignora instâncias que estão protegidas contra a redução da escala horizontalmente e continua substituindo instâncias que não estão protegidas.
- Aguardar (`Wait`): aguarda uma hora até você remover a proteção contra redução da escala na horizontal. Se você não fizer isso, a atualização de instância falhará.

Ignorar correspondência (**`SkipMatching`**)

Controla se o Amazon EC2 Auto Scaling ignora a substituição de instâncias que correspondam à configuração desejada. Se nenhuma configuração desejada for especificada, ele ignorará a substituição de instâncias que tenham o mesmo modelo de execução e tipos de instância que o grupo do Auto Scaling estava usando antes do início da atualização de instância. Para obter mais informações, consulte Usar uma atualização de instância com opção de ignorar correspondência.

Instâncias em espera (**`StandbyInstances`**)

Controla o que o Amazon EC2 Auto Scaling faz se as instâncias forem encontradas no estado`Standby`. Para obter mais informações sobre essas instâncias, consulte Remover temporiamente instâncias do grupo do Auto Scaling.

O Amazon EC2 Auto Scaling fornece estas opções:

- Terminar (`Terminate`): termina instâncias que estejam em `Standby`.
- Ignorar (`Ignore`): ignora instâncias que estejam em `Standby` e continua substituindo instâncias que estejam no estado `InService`.
- Aguardar (**`Wait`**): aguarda uma hora para você retornar as instâncias ao serviço. Se você não fizer isso, a atualização de instância falhará.

Inicie uma atualização de instância usando o ou Console de gerenciamento da AWS AWS CLI

> **Important**
>
> É possível reverter uma atualização de instância que esteja em andamento para desfazer alterações. Para que isso funcione, o grupo do Auto Scaling deve atender aos pré-requisitos para uso de reversões antes de iniciar a atualização de instância. Para obter mais informações, consulte Desfazer alterações com uma reversão manual ou automática.

Os procedimentos a seguir ajudam você a iniciar uma atualização de instância usando o Console de gerenciamento da AWS ou AWS CLI.

Iniciar uma atualização de instância (console)

Se esta for a primeira vez que inicia uma atualização de instância, fazer isso usando o console ajudará você a entender os recursos e as opções disponíveis.

Iniciar uma atualização de instância no console (procedimento básico)

Use o procedimento a seguir se você não tiver definido anteriormente uma política de instâncias mistas para seu grupo do Auto Scaling. Se você já definiu uma política de instâncias mistas, consulte Iniciar uma atualização de instância no console (grupo de instâncias mistas) para iniciar uma atualização de instância.

Para iniciar uma atualização de instância

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Atualização de instância, em Atualização ativa de instância, escolha Iniciar atualização de instância.
4. Para configurações de disponibilidade, faça o seguinte:

   a. Para o método de substituição de instância:

      - Se você não definiu uma política de manutenção de instâncias no grupo do Auto Scaling, a configuração padrão para o método de substituição de instância é Encerrar e iniciar. Esse é o comportamento padrão legado de uma atualização de instância.
      - Se você definir uma política de manutenção de instância no grupo do Auto Scaling, ela fornecerá valores padrão para o método de substituição de instância. Para substituir a política de manutenção da instância, escolha Substituir. A substituição é aplicada somente à atualização de instância atual. Na próxima vez que você iniciar uma atualização de instância, esses valores serão redefinidos para os padrões da política de manutenção de instâncias.

      O procedimento a seguir explica como atualizar o método de substituição de instância.

      i. Escolha um dos seguintes métodos de substituição de instância:

         - Iniciar antes de encerrar: uma nova instância deve ser provisionada primeiro antes que uma instância existente possa ser encerrada. Essa é uma boa opção para aplicações que favorecem a disponibilidade em detrimento da redução de custos.
         - Encerrar e iniciar: novas instâncias são provisionadas ao mesmo tempo em que as instâncias existentes são encerradas. Esta é uma boa opção para aplicações que favorecem a economia de custos em detrimento da disponibilidade. Também é uma boa opção para aplicações que não devem lançar mais capacidade do que a disponível atualmente.

- Comportamento personalizado: esta opção permite configurar um intervalo mínimo e máximo personalizado para a quantidade de capacidade que você deseja disponibilizar ao substituir instâncias. Isso pode ajudar a alcançar o equilíbrio certo entre custo e disponibilidade.

ii. Em Definir porcentagem de integridade, insira valores para um ou ambos os campos a seguir. Os campos de ativação variam de acordo com a opção escolhida para o método de substituição de instância.

- Mínimo: define a porcentagem mínima de integridade necessária para continuar com a atualização de instâncias.
- Máximo: Define a porcentagem máxima íntegra possível durante a atualização da instância.

iii. Expanda a seção Exibir capacidade temporária estimada durante substituições com base no tamanho atual do grupo para confirmar como os valores de Mínimo e Máximo se aplicam ao seu grupo. Os valores exatos usados dependem do valor de capacidade desejado, que mudará se o grupo for ampliado.

iv. Expanda a seção Definir comportamento alternativo para tamanhos de reposição inválidos e, em seguida, escolha se deseja violar a porcentagem máxima de integridade para priorizar a disponibilidade ou violar a porcentagem mínima de integridade.

Manter a opção padrão de Violar porcentagem mínima de integridade não é recomendado para grupos muito pequenos. Se houver apenas uma instância no grupo do Auto Scaling, iniciar uma atualização de instância poderá resultar em uma interrupção.

Essa etapa configura o comportamento de fallback se você estiver usando um grupo do Auto Scaling que ainda não tem uma política de manutenção de instâncias. Essa opção não está disponível e não aparece quando seu grupo tem uma política de manutenção de instâncias. Essa opção também está disponível somente para o método de substituição Encerrar e iniciar. Outros métodos de substituição violarão a porcentagem máxima de integridade para priorizar a disponibilidade.

b. Em Aquecimento da instância, insira o número de segundos desde a mudança do estado de uma nova instância até o `InService` término da inicialização. O Amazon EC2 Auto Scaling aguarda esse tempo antes de substituir a próxima instância.

Durante o aquecimento, instâncias recém-iniciadas também não são contabilizadas nas métricas agregadas do grupo do Auto Scaling (como `CPUUtilization`, `NetworkIn`, `NetworkOut` etc.). Se você adicionou políticas de escalabilidade ao grupo do Auto Scaling, as ações de escalabilidade serão executadas em paralelo. Se você definir um intervalo longo para o período de aquecimento de atualização de instância, levará mais tempo para que as instâncias recém-iniciadas sejam exibidas nas métricas. Portanto, um período de aquecimento adequado evita a escalabilidade do Amazon EC2 Auto Scaling em dados de métricas obsoletos.

Se você já definiu corretamente um aquecimento de instâncias padrão para o grupo do Auto Scaling, não é necessário alterar o aquecimento da instância. Porém, se quiser substituir o padrão, você pode definir um valor para essa opção. Para obter mais informações sobre como configurar o aquecimento de instâncias, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

5. Para configurações de atualização, faça o seguinte:

   a. (Opcional) Em Pontos de verificação, escolha Habilitar pontos de verificação para substituir instâncias usando uma abordagem incremental ou faseada para uma atualização de instância. Isso fornece tempo adicional para verificação entre conjuntos de substituições. Se você optar por não ativar pontos de verificação, as instâncias serão substituídas em uma operação quase contínua.

      Se você habilitar pontos de verificação, consulte Habilitar pontos de verificação (console) para obter etapas adicionais.

   b. (Opcional) Em Tempo de incorporação especifique a quantidade de tempo de espera no final de uma atualização de instância antes que a atualização da instância seja considerada concluída.

   c. Habilitar ou desativar Ignorar correspondência :

      - Para ignorar a substituição de instâncias que já correspondem ao modelo de execução, mantenha a caixa de seleção Habilitar opção de ignorar correspondência marcada.
      - Se você desativar ignorar correspondência desmarcando essa caixa de seleção, todas as instâncias poderão ser substituídas.

Ao ativar a correspondência ignorada, você pode definir um novo modelo de execução ou uma nova versão do modelo de execução em vez de usar o existente. Faça isso na seção Configuração desejada da página Iniciar atualização de instância.

> **Note**
>
> Para usar o recurso de ignorar correspondência para atualizar um grupo do Auto Scaling que atualmente use uma configuração de execução, é necessário selecionar um modelo de execução em Configuração desejada. Ignorar correspondência com uma configuração de inicialização não é compatível.

d. Em Instâncias em espera, escolha Ignorar, Terminar ou Aguardar. Isso determina o que acontecerá se as instâncias forem encontradas no estado `Standby`. Para obter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.

   Se você escolher Aguardar, deverá realizar outras ações para retornar essas instâncias ao serviço. Senão, a atualização de instância substituirá todas as instâncias `InService` e aguardará uma hora. Então, se alguma instância `Standby` permanecer, a atualização de instância falhará. Para evitar essa situação, escolha Ignorar ou Terminar as instâncias.

e. Para Instâncias protegidas de redução da escala na horizontal, escolha Ignorar, Substituir ou Aguardar. Isso determina o que acontecerá se instâncias protegidas contra redução da escala na horizontal forem encontradas. Para obter mais informações, consulte Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância.

   Se você escolher Aguardar, deverá realizar outras ações para remover a proteção contra redução da escala na horizontal dessas instâncias. Senão, a atualização de instância substituirá todas as instâncias não protegidas e aguardará uma hora. Então, se alguma instância protegida contra redução da escala na horizontal permanecer, a atualização de instância falhará. Para evitar essa situação, escolha Ignorar ou Substituir as instâncias.

6. (Opcional) Para CloudWatch alarme, escolha Ativar CloudWatch alarmes e, em seguida, escolha um ou mais alarmes. CloudWatch os alarmes podem ser usados para identificar quaisquer problemas e falhar na operação se um alarme entrar no `ALARM` estado. Para obter mais informações, consulte Iniciar uma atualização de instância com reversão automática.
7. (Opcional) Expanda a seção Configuração desejada para especificar as atualizações que você deseja fazer no grupo do Auto Scaling.

Nesta etapa, você pode optar por usar a sintaxe JSON ou YAML para editar valores de parâmetros em vez de fazer seleções na interface do console. Para isso, escolha Usar editor de código em vez de Usar a interface do console. O procedimento a seguir explica como fazer seleções usando a interface do console.

a. Para Atualizar o modelo de execução:

   - Se você não criou um novo modelo de execução ou uma nova versão de modelo de execução para seu grupo do Auto Scaling, não marque essa caixa de seleção.
   - Se você criou um novo modelo de execução ou uma nova versão do modelo de execução, marque esta caixa de seleção. Quando você seleciona essa opção, o Amazon EC2 Auto Scaling exibe o modelo de execução atual e a versão atual do modelo de execução. Também lista todas as outras versões disponíveis. Escolha o modelo de lançamento e, em seguida, escolha a versão.

     Após escolher uma versão, você poderá visualizar as informações da versão. Esta é a versão do modelo de execução que será usada ao substituir instâncias como parte de uma atualização de instância. Se a atualização da instância tiver êxito, essa versão do modelo de execução também será usada sempre que novas instâncias forem iniciadas, como quando o grupo for dimensionado.

b. Em Choose a set of instance types and purchase options to override the instance type in the launch template (Escolha um conjunto de tipos de instância e opções de compra para substituir o tipo de instância no modelo de execução):

   - Não marque essa caixa de seleção se quiser usar o tipo de instância e a opção de compra que você especificou no modelo de execução.
   - Marque esta caixa de seleção se quiser substituir o tipo de instância no modelo de execução ou executar instâncias spot. É possível adicionar manualmente cada tipo de instância ou escolher um tipo de instância primária e uma opção de recomendação que recupere outros tipos de instância correspondentes para você. Se você pretende iniciar instâncias spot, recomendamos adicionar alguns tipos diferentes de instância. Dessa forma, o Amazon EC2 Auto Scaling pode iniciar outro tipo de instância se houver capacidade de instância insuficiente nas zonas de disponibilidade escolhidas. Para obter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

> **Warning**
>
> Não use instâncias spot com aplicações que não conseguem lidar com uma interrupção de instância spot. As interrupções poderão ocorrer se o serviço do Amazon EC2 Spot precisar recuperar a capacidade.

Se você marcar essa caixa de seleção, verifique se o modelo de execução já não solicita instâncias spot. Não é possível usar um modelo de execução que solicite instâncias spot para criar um grupo do Auto Scaling que use vários tipos de instância e execute instâncias spot e sob demanda.

> **Note**
>
> Para configurar essas opções em um grupo do Auto Scaling que atualmente use uma configuração de execução, é necessário selecionar um modelo de execução em Update launch template (Atualizar modelo de execução). Não há suporte à substituição do tipo de instância na configuração de execução.

8. (Opcional) Em Configurações de reversão, escolha Habilitar reversão automática para reverter automaticamente a atualização de instância em caso de falha.

   Essa configuração só pode ser habilitada quando o grupo do Auto Scaling atende aos pré-requisitos para usar reversões.

   Para obter mais informações, consulte Desfazer alterações com uma reversão manual ou automática.

9. Revise todas as seleções para confirmar que tudo esteja configurado corretamente.

   Nesse ponto, é bom verificar se as diferenças entre as alterações atuais e propostas não afetarão sua aplicação de maneiras inesperadas ou indesejadas. Para confirmar se o tipo de instância é compatível com o modelo de execução, consulte Compatibilidade de tipo de instância.

10. Quando estiver satisfeito com as seleções de atualização da instância, escolha Iniciar atualização da instância.

Iniciar uma atualização de instância no console (grupo de instâncias mistas)

Use o procedimento a seguir se você criou um grupo do Auto Scaling com política de instâncias mistas. Se você não definiu ainda uma política de instâncias mistas para seu grupo, consulte Iniciar uma atualização de instância no console (procedimento básico) para iniciar uma atualização de instância.

Para iniciar uma atualização de instância

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Atualização de instância, em Atualização ativa de instância, escolha Iniciar atualização de instância.
4. Para configurações de disponibilidade, faça o seguinte:

   a. Para o método de substituição de instância:

      - Se você não definiu uma política de manutenção de instâncias no grupo do Auto Scaling, a configuração padrão para o método de substituição de instância é Encerrar e iniciar. Esse é o comportamento padrão legado de uma atualização de instância.
      - Se você definir uma política de manutenção de instância no grupo do Auto Scaling, ela fornecerá valores padrão para o método de substituição de instância. Para substituir a política de manutenção da instância, escolha Substituir. A substituição é aplicada somente à atualização de instância atual. Na próxima vez que você iniciar uma atualização de instância, esses valores serão redefinidos para os padrões da política de manutenção de instâncias.

      O procedimento a seguir explica como atualizar o método de substituição de instância.

      i. Escolha um dos seguintes métodos de substituição de instância:

         - Iniciar antes de encerrar: uma nova instância deve ser provisionada primeiro antes que uma instância existente possa ser encerrada. Essa é uma boa opção para aplicações que favorecem a disponibilidade em detrimento da redução de custos.

- Encerrar e iniciar: novas instâncias são provisionadas ao mesmo tempo em que as instâncias existentes são encerradas. Esta é uma boa opção para aplicações que favorecem a economia de custos em detrimento da disponibilidade. Também é uma boa opção para aplicações que não devem lançar mais capacidade do que a disponível atualmente.
- Comportamento personalizado: esta opção permite configurar um intervalo mínimo e máximo personalizado para a quantidade de capacidade que você deseja disponibilizar ao substituir instâncias. Isso pode ajudar a alcançar o equilíbrio certo entre custo e disponibilidade.

ii. Em Definir porcentagem de integridade, insira valores para um ou ambos os campos a seguir. Os campos de ativação variam de acordo com a opção escolhida para o método de substituição de instância.

- Mínimo: define a porcentagem mínima de integridade necessária para continuar com a atualização de instâncias.
- Máximo: Define a porcentagem máxima íntegra possível durante a atualização da instância.

iii. Expanda a seção Exibir capacidade temporária estimada durante substituições com base no tamanho atual do grupo para confirmar como os valores de Mínimo e Máximo se aplicam ao seu grupo. Os valores exatos usados dependem do valor de capacidade desejado, que mudará se o grupo for ampliado.

iv. Expanda a seção Definir comportamento alternativo para tamanhos de reposição inválidos e, em seguida, escolha se deseja violar a porcentagem máxima de integridade para priorizar a disponibilidade ou violar a porcentagem mínima de integridade.

Manter a opção padrão de Violar porcentagem mínima de integridade não é recomendado para grupos muito pequenos. Se houver apenas uma instância no grupo do Auto Scaling, iniciar uma atualização de instância poderá resultar em uma interrupção.

Essa etapa configura o comportamento de fallback se você estiver usando um grupo do Auto Scaling que ainda não tem uma política de manutenção de instâncias. Essa opção não está disponível e não aparece quando seu grupo tem uma política de manutenção de instâncias. Essa opção também está disponível somente para o

método de substituição Encerrar e iniciar. Outros métodos de substituição violarão a porcentagem máxima de integridade para priorizar a disponibilidade.

b. Em Aquecimento da instância, insira o número de segundos desde a mudança do estado de uma nova instância até o `InService` término da inicialização. O Amazon EC2 Auto Scaling aguarda esse tempo antes de substituir a próxima instância.

   Durante o aquecimento, instâncias recém-iniciadas também não são contabilizadas nas métricas agregadas do grupo do Auto Scaling (como `CPUUtilization`, `NetworkIn`, `NetworkOut` etc.). Se você adicionou políticas de escalabilidade ao grupo do Auto Scaling, as ações de escalabilidade serão executadas em paralelo. Se você definir um intervalo longo para o período de aquecimento de atualização de instância, levará mais tempo para que as instâncias recém-iniciadas sejam exibidas nas métricas. Portanto, um período de aquecimento adequado evita a escalabilidade do Amazon EC2 Auto Scaling em dados de métricas obsoletos.

   Se você já definiu corretamente um aquecimento de instâncias padrão para o grupo do Auto Scaling, não é necessário alterar o aquecimento da instância. Porém, se quiser substituir o padrão, você pode definir um valor para essa opção. Para obter mais informações sobre como configurar o aquecimento de instâncias, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

5. Para configurações de atualização, faça o seguinte:

   a. (Opcional) Em Pontos de verificação, escolha Habilitar pontos de verificação para substituir instâncias usando uma abordagem incremental ou faseada para uma atualização de instância. Isso fornece tempo adicional para verificação entre conjuntos de substituições. Se você optar por não ativar pontos de verificação, as instâncias serão substituídas em uma operação quase contínua.

      Se você habilitar pontos de verificação, consulte Habilitar pontos de verificação (console) para obter etapas adicionais.

   b. Habilitar ou desativar Ignorar correspondência:

      - Para ignorar a substituição de instâncias que já correspondem ao modelo de execução e quaisquer substituições de tipo de instância, mantenha a caixa de seleção Habilitar opção de ignorar correspondência marcada.
      - Se você optar por desativar ignorar correspondência desmarcando essa caixa de seleção, todas as instâncias poderão ser substituídas.

Ao ativar a correspondência ignorada, você pode definir um novo modelo de execução ou uma nova versão do modelo de execução em vez de usar o existente. Faça isso na seção Configuração desejada da página Iniciar atualização de instância. Você também pode atualizar suas substituições de tipo de instância em Desired configuration (Configuração desejada).

c. Em Instâncias em espera, escolha Ignorar, Terminar ou Aguardar. Isso determina o que acontecerá se as instâncias forem encontradas no estado `Standby`. Para obter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.

   Se você escolher Aguardar, deverá realizar outras ações para retornar essas instâncias ao serviço. Do contrário, a atualização de instância substituirá todas as instâncias `InService` e aguardará uma hora. Então, se alguma instância `Standby` permanecer, a atualização de instância falhará. Para evitar essa situação, escolha Ignorar ou Terminar as instâncias.

d. Para Instâncias protegidas de redução da escala na horizontal, escolha Ignorar, Substituir ou Aguardar. Isso determina o que acontecerá se instâncias protegidas contra redução da escala na horizontal forem encontradas. Para obter mais informações, consulte Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância.

   Se você escolher Aguardar, deverá realizar outras ações para remover a proteção contra redução da escala na horizontal dessas instâncias. Senão, a atualização de instância substituirá todas as instâncias não protegidas e aguardará uma hora. Então, se alguma instância protegida contra redução da escala na horizontal permanecer, a atualização de instância falhará. Para evitar essa situação, escolha Ignorar ou Substituir as instâncias.

6. (Opcional) Para CloudWatch alarme, escolha Ativar CloudWatch alarmes e, em seguida, escolha um ou mais alarmes. CloudWatch os alarmes podem ser usados para identificar quaisquer problemas e falhar na operação se um alarme entrar no `ALARM` estado. Para obter mais informações, consulte Iniciar uma atualização de instância com reversão automática.

7. Na seção Desired configuration (Configuração desejada), faça o seguinte:

   Nesta etapa, você pode optar por usar a sintaxe JSON ou YAML para editar valores de parâmetros em vez de fazer seleções na interface do console. Para isso, escolha Usar editor de código em vez de Usar a interface do console. O procedimento a seguir explica como fazer seleções usando a interface do console.

   a. Para Atualizar o modelo de execução:

- Se você não criou um novo modelo de execução ou uma nova versão de modelo de execução para seu grupo do Auto Scaling, não marque essa caixa de seleção.
- Se você criou um novo modelo de execução ou uma nova versão do modelo de execução, marque esta caixa de seleção. Quando você seleciona essa opção, o Amazon EC2 Auto Scaling exibe o modelo de execução atual e a versão atual do modelo de execução. Também lista todas as outras versões disponíveis. Escolha o modelo de lançamento e, em seguida, escolha a versão.

  Após escolher uma versão, você poderá visualizar as informações da versão. Esta é a versão do modelo de execução que será usada ao substituir instâncias como parte de uma atualização de instância. Se a atualização da instância tiver êxito, essa versão do modelo de execução também será usada sempre que novas instâncias forem iniciadas, como quando o grupo for dimensionado.

b. Em Use these settings to override the instance type and purchase option defined in the launch template (Use estas configurações para substituir o tipo de instância e a opção de compra definidas no modelo de execução):

   Por padrão, esta caixa de seleção está marcada. O Amazon EC2 Auto Scaling preenche cada parâmetro com o valor que está atualmente definido na política de instâncias mistas para o grupo do Auto Scaling. Atualize somente os valores dos parâmetros que você deseja alterar. Para obter orientações sobre essas configurações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

   > ⚠ Warning
   >
   > Recomendamos não desmarcar essa caixa de seleção. Apenas a desmarque se desejar parar de usar uma política de instâncias mistas. Após o término com êxito da atualização de instância, o Amazon EC2 Auto Scaling atualiza seu grupo para corresponder à Desired configuration (Configuração desejada). Se não incluir mais uma política de instâncias mistas, o Amazon EC2 Auto Scaling terminará gradualmente todas as instâncias spot que estejam em execução no momento e as substituirá por instâncias sob demanda. Ou, se seu modelo de execução solicitar instâncias spot, o Amazon EC2 Auto Scaling terminará gradualmente todas as instâncias sob demanda que estejam em execução no momento e as substituirá por instâncias spot.

8. (Opcional) Em Configurações de reversão, escolha Habilitar reversão automática para reverter automaticamente a atualização de instância em caso de falha por qualquer motivo.

   Essa configuração só pode ser habilitada quando o grupo do Auto Scaling atende aos pré-requisitos para usar reversões.

   Para obter mais informações, consulte Desfazer alterações com uma reversão manual ou automática.
9. Revise todas as seleções para confirmar que tudo esteja configurado corretamente.

   Nesse ponto, é bom verificar se as diferenças entre as alterações atuais e propostas não afetarão sua aplicação de maneiras inesperadas ou indesejadas. Para confirmar se o tipo de instância é compatível com o modelo de execução, consulte Compatibilidade de tipo de instância.

   Quando estiver satisfeito com as seleções de atualização da instância, escolha Iniciar atualização da instância.

Iniciar uma atualização de instância (AWS CLI)

Para iniciar uma atualização de instância

Use o start-instance-refreshcomando a seguir para iniciar uma atualização de instância a AWS CLI partir do. Você pode especificar as preferências que deseja alterar em um arquivo de configuração JSON. Ao referenciar o arquivo de configuração, forneça o caminho e o nome do arquivo, conforme mostrado no exemplo a seguir.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 50,
      "AutoRollback": true,
      "ScaleInProtectedInstances": Ignore,
      "StandbyInstances": Terminate
    }
```

```
}
```

Se as preferências não forem fornecidas, serão usados os valores padrão. Para obter mais informações, consulte Entender os valores padrão de uma atualização de instância.

Resultado do exemplo:

```
{
    "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

Monitore a atualização de uma instância usando o ou Console de gerenciamento da AWS AWS CLI

Você pode monitorar uma atualização de instância em andamento ou consultar o status de atualizações de instâncias anteriores nas últimas seis semanas usando o Console de gerenciamento da AWS ou. AWS CLI

Para verificar o status de uma atualização de instância

Para relatar o status da instância, use um dos seguintes métodos:

Console

> **Tip**
>
> Em cada uma delas, as colunas nomeadas já deverão ser exibidas. Para exibir colunas ocultas ou alterar o número de linhas exibidas, escolha o ícone de engrenagem no canto superior direito de cada seção para abrir o modal de preferências. Atualize as configurações, conforme necessário, e escolha Confirmar.

Para verificar o status de uma atualização de instância (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.

3. Na guia Instance refresh (Atualização da instância), em Instance refresh history (Histórico da atualização de instâncias), é possível determinar o status da sua solicitação observando a coluna Status. A operação entra no status `Pending` durante a inicialização. Depois, o status deve mudar rapidamente para `InProgress`. Quando todas as instâncias estão atualizadas, o status muda para `Successful`.
4. Você pode monitorar ainda mais o sucesso ou o fracasso das atividades em andamento, visualizando as atividades de escala do grupo. Na guia Atividade, em Histórico de atividades, quando a atualização da instância for iniciada, você observará entradas quando instâncias forem encerradas e outro conjunto de entradas quando instâncias forem iniciadas. (Opcional) Caso tenha muitas atividades de escalabilidade, você poderá ver mais delas escolhendo o ícone **>** na parte superior do histórico de atividades. Para obter informações sobre a solução de problemas que podem causar falhas nas atividades, consulte Solucionar problemas do Amazon EC2 Auto Scaling.
5. (Opcional) Na guia Gerenciamento de instâncias, em Instâncias, você pode revisar o progresso de instâncias específicas, conforme necessário.

AWS CLI

Como verificar o status de uma atualização de instância (AWS CLI)

Use o seguinte comando describe-instance-refreshes:

```
aws autoscaling describe-instance-refreshes --auto-scaling-group-name my-asg
```

O seguinte é um exemplo de saída.

As atualizações de instância são ordenadas pela hora de início. As atualizações de instância ainda em andamento são descritas primeiro.

```
{
  "InstanceRefreshes":[
    {
      "InstanceRefreshId":"08b91cf7-8fa6-48af-b6a6-d227f40f1b9b",
      "AutoScalingGroupName":"my-asg",
      "Status":"InProgress",
      "StatusReason":"Waiting for instances to warm up before continuing. For
 example: i-0645704820a8e83ff is warming up.",
      "StartTime":"2023-11-24T16:46:52+00:00",
```

```
      "PercentageComplete":50,
      "InstancesToUpdate":0,
      "Preferences":{
        "MaxHealthyPercentage":120,
        "MinHealthyPercentage":90,
        "InstanceWarmup":60,
        "SkipMatching":false,
        "AutoRollback":true,
        "ScaleInProtectedInstances":"Ignore",
        "StandbyInstances":"Ignore"
      }
    },
    {
      "InstanceRefreshId":"0e151305-1e57-4a32-a256-1fd14157c5ec",
      "AutoScalingGroupName":"my-asg",
      "Status":"Successful",
      "StartTime":"2023-11-22T13:53:37+00:00",
      "EndTime":"2023-11-22T13:59:45+00:00",
      "PercentageComplete":100,
      "InstancesToUpdate":0,
      "Preferences":{
        "MaxHealthyPercentage":120,
        "MinHealthyPercentage":90,
        "InstanceWarmup":60,
        "SkipMatching":false,
        "AutoRollback":true,
        "ScaleInProtectedInstances":"Ignore",
        "StandbyInstances":"Ignore"
      }
    }
  ]
}
```


Você pode monitorar ainda mais o sucesso ou o fracasso das atividades em andamento, visualizando as atividades de escala do grupo. As atividades de escala também o ajudam a obter mais detalhes para solucionar problemas com a atualização de uma instância. Para obter mais informações, consulte Solucionar problemas do Amazon EC2 Auto Scaling.

Status de atualização de instância

Quando uma atualização de instância é iniciada, ela entra no status Pending. Ele passa de Pendente para InProgressaté atingir Sucesso, Falha, Cancelado ou RollbackFailed. RollbackSuccessful

A atualização de instância pode ter os seguintes status:

Status	Description
Pendente	A solicitação foi criada, mas a atualização de instância não foi iniciada.
InProgress	Uma atualização de instância está em andamento.
Com êxito	Uma atualização de instância foi concluída com êxito.
Com falha	Falha ao concluir uma atualização de instância. É possível solucionar problemas usando o motivo do status e as ações de escalabilidade.
Cancelando	Uma atualização de instância em andamento está sendo cancelada.
Cancelado	A atualização de instância foi cancelada.
RollbackInProgress	Uma atualização de instância está sendo revertida.
RollbackFailed	Falha ao concluir a reversão. É possível solucionar problemas usando o motivo do status e as ações de escalabilidade.
RollbackSuccessful	A reversão foi concluída com êxito.
Incorporação	Aguardando o tempo de incorporação especificado após a conclusão da atualização de uma instância.

Substitua os volumes raiz durante a atualização da instância

Conteúdo

- Como funcionam
- Requisitos
- Inicie uma atualização de instância com Replace Root Volume
- Limitações
- Use ganchos de ciclo de vida para substituir o volume raiz

Como funcionam

A substituição do volume raiz atualiza suas instâncias substituindo somente o volume raiz do EBS enquanto mantém a instância em execução. Isso elimina a necessidade de lançar novas instâncias e evita possíveis restrições de capacidade. Esse processo preserva todos os outros recursos da instância, incluindo:

- Interfaces de rede e endereços IP
- Volumes não raiz do EBS
- Volumes e dados do armazenamento de instâncias
- Grupos de segurança e funções do IAM

Durante a substituição do volume raiz, seu aplicativo continua em execução na instância existente. O volume raiz original é separado, um novo volume raiz é criado a partir da AMI especificada e, em seguida, anexado à mesma instância. Os volumes raiz originais são excluídos automaticamente após a substituição bem-sucedida.

Quando você usa a estratégia de substituição do volume raiz, as instâncias passam pelos seguintes estados do ciclo de vida:

1. `ReplacingRootVolume`- A substituição do volume raiz começa
2. `ReplacingRootVolume:Wait`- Aguarda a conclusão do ciclo de vida do gancho (se configurado)
3. `ReplacingRootVolume:Proceed`- Prossegue com a substituição
4. `RootVolumeReplaced`- Substituição concluída com sucesso

Após a conclusão da substituição do volume raiz, o Auto Scaling realiza verificações de integridade nas instâncias atualizadas. Se a substituição do volume raiz falhar em uma instância, o Auto Scaling marca a instância para encerramento e a substitui por uma nova instância.

Requisitos

- Seu grupo de Auto Scaling deve usar uma política de instâncias mistas
- Todas as substituições na política de instâncias mistas devem especificar um `ImageId`
- AMIs deve conter somente um único volume raiz
- Todas as instâncias devem corresponder à configuração do modelo de lançamento do grupo

- Você deve iniciar a atualização da instância com a configuração desejada, com uma política de instâncias mistas contendo `ImageId` substituições.

Inicie uma atualização de instância com Replace Root Volume

Para iniciar uma atualização de instância com Replace Root Volume ()AWS CLI

Use o start-instance-refreshcomando a seguir para iniciar uma atualização de instância a AWS CLI partir do. Você pode especificar as preferências que deseja alterar em um arquivo de configuração JSON. Ao referenciar o arquivo de configuração, forneça o caminho e o nome do arquivo, conforme mostrado no exemplo a seguir.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`:

Example

```
{
  "AutoScalingGroupName": "my-asg",
  "Strategy" : "ReplaceRootVolume",
  "DesiredConfiguration":{
    "MixedInstancesPolicy":{
      "LaunchTemplate": {
        "LaunchTemplateSpecification": {
          "LaunchTemplateId":"my-launch-template",
          "Version": "1"
        },
        "Overrides": [
          {
            "InstanceType":"c5.large",
            "ImageId":"ami-1234example"
          },
          {
            "InstanceType":"m5.large",
            "ImageId":"ami-2345example"
          }
        ]
      }
    }
  },
```

```
  "Preferences": {
    "InstanceWarmup": 60,
    "MinHealthyPercentage": 90,
    "AutoRollback": true,
    "ScaleInProtectedInstances": Ignore,
    "StandbyInstances": Ignore
  }
}
```

Se as preferências não forem fornecidas, serão usados os valores padrão. Para obter mais informações, consulte Entender os valores padrão de uma atualização de instância.

Resultado do exemplo:

```
{
    "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

Limitações

- Você não pode iniciar uma atualização de instância com Replace Root Volume se o grupo EC2 Auto Scaling ou a configuração desejada de atualização de instância usar `$Latest` a versão ou executar o modelo. `$Default`
- Você não pode iniciar uma atualização de instância com Replace Root Volume em um grupo do Amazon EC2 Auto Scaling se ele contiver instâncias cujo tipo de instância não esteja presente na política de instâncias mistas.
- Durante uma atualização de instância com Replace Root Volume, você só pode anexar instâncias com tipos que estejam presentes na política de instâncias mistas da configuração desejada.
- Você não pode iniciar uma atualização de instância com Replace Root Volume em um grupo de EC2 Auto Scaling que tenha um pool aquecido.
- Você não pode adicionar um pool aquecido a um grupo do Amazon EC2 Auto Scaling que tenha uma atualização de instância ativa com Replace Root Volume.

Use ganchos de ciclo de vida para substituir o volume raiz

As instâncias submetidas à substituição do volume raiz seguem suas próprias transições de ciclo de vida, permitindo que você invoque ações (por exemplo, uma função Lambda) antes da substituição.

and/or Para obter mais informações sobre os estados do ciclo de vida durante a substituição do volume raiz, consulte. Transições de estado do ciclo de vida para instâncias que estão passando por substituição do volume raiz

Para obter mais informações sobre a adição de um gancho do ciclo de vida, consulte Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling. Para obter mais informações sobre a conclusão de uma ação do ciclo de vida, consulte Concluir uma ação do ciclo de vida em um grupo do Auto Scaling.

Antes da substituição do volume raiz, um gancho de ciclo de vida pode ser útil nos seguintes cenários:

- Você deseja encerrar seu aplicativo normalmente antes que o volume raiz seja substituído e a instância seja reinicializada.
- Você deseja mover os dados do volume raiz para outro local antes que o volume seja substituído.

Após a substituição do volume raiz, um gancho de ciclo de vida pode ser útil nos seguintes cenários:

- Você quer verificar se sua instância está totalmente pronta com a nova AMI antes de começar a receber tráfego.
- Você deseja inicializar os dados do aplicativo que existiam anteriormente no volume raiz.

Ao adicionar ganchos do ciclo de vida, considere o seguinte:

- Quando um gancho de ciclo de vida é configurado para a ação do `autoscaling:EC2_INSTANCE_TERMINATING` ciclo de vida, uma instância prestes a passar pela substituição do volume raiz faz uma pausa para realizar uma ação personalizada quando atinge o estado. `ReplacingRootVolume:Wait`
- Quando um gancho de ciclo de vida é configurado para a ação do `autoscaling:EC2_INSTANCE_LAUNCHING` ciclo de vida, uma instância que acabou de ter seu volume raiz substituído faz uma pausa para realizar uma ação personalizada quando atinge o estado. `Pending:Wait`

Quando as instâncias atingem um estado de espera, o Amazon EC2 Auto Scaling envia uma notificação. As notificações de gancho do ciclo de vida do Auto Scaling contêm um `Action` campo com valor `ReplaceRootVolume` para indicar que a instância está sendo substituída pelo volume raiz.

Exemplos dessas notificações estão disponíveis na EventBridge seção deste guia. Para obter mais informações, consulte A atualização da instância substitui os eventos do ciclo de vida do volume raiz.

Cancelar uma atualização de instância usando o ou Console de gerenciamento da AWS AWS CLI

É possível cancelar uma atualização de instância que ainda esteja em andamento. Não é possível cancelá-la após a conclusão.

Cancelar uma atualização de instância não reverterá instâncias que já foram substituídas. Em vez disso, para reverter as alterações das instâncias, realize uma reversão. Para obter mais informações, consulte Desfazer alterações com uma reversão manual ou automática.

Tópicos

- Cancelar uma atualização de instância (console)
- Cancelar uma atualização de instância (AWS CLI)

Cancelar uma atualização de instância (console)

Como cancelar uma atualização de instância

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.
3. Na guia Atualização de instância, em Atualização de instância ativa, escolha Ações, Cancelar.
4. Quando a confirmação for solicitada, escolha Confirmar.

O status da atualização de instância está definido como Cancelling. Depois que o cancelamento for concluído, o status da atualização de instância será definido como Cancelled.

Cancelar uma atualização de instância (AWS CLI)

Como cancelar uma atualização de instância

Use o cancel-instance-refreshcomando do AWS CLI e forneça o nome do grupo Auto Scaling.

```
aws autoscaling cancel-instance-refresh --auto-scaling-group-name my-asg
```

Resultado do exemplo:

```
{
    "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

Como cancelar uma atualização de instância sem esperar pela transição da instância

Use a `--no-wait-for-transitioning-instances` opção com o cancel-instance-refreshcomando do AWS CLI para cancelar a atualização de uma instância sem esperar pela transição das instâncias.

```
aws autoscaling cancel-instance-refresh –auto-scaling-group-name my-asg --no-wait-for-transitioning-instances
```

Por padrão, quando você cancela uma atualização de instância, o Amazon EC2 Auto Scaling aguarda por todas as execuções e encerramentos em andamento antes de definir o status como Cancelado. Se você tiver instâncias com longos períodos de drenagem ou precisar iniciar uma nova atualização de instância imediatamente, poderá usar o comando para não precisar esperar por essas operações de transição.

Quando você usa essa opção, o status de atualização da instância muda para Cancelado, permitindo que você inicie uma nova atualização da instância. Todas as execuções e encerramentos em andamento prosseguem em segundo plano.

Desfazer alterações com uma reversão manual ou automática

É possível reverter uma atualização de instância que ainda esteja em andamento. Não é possível revertê-la após a conclusão. Porém, você pode atualizar seu grupo do Auto Scaling novamente iniciando uma nova atualização de instância.

Durante a reversão, o Amazon EC2 Auto Scaling substitui as instâncias que foram implantadas até o momento. As novas instâncias correspondem à configuração que você salvou pela última vez no grupo do Auto Scaling antes de iniciar a atualização de instância.

O Amazon EC2 Auto Scaling fornece estes modos de reversão:

- Reversão manual: inicie uma reversão manualmente para reverter o que foi implantado até o ponto de reversão.

- Reversão automática: o Amazon EC2 Auto Scaling reverte automaticamente o que foi implantado se a atualização da instância falhar por algum motivo ou se CloudWatch algum alarme especificado entrar no estado. `ALARM`

Conteúdo

- Considerações
- Iniciar uma reversão manualmente
- Iniciar uma atualização de instância com reversão automática

Considerações

As seguintes considerações se aplicam ao usar uma reversão:

- A opção de reversão só está disponível se você especificar a configuração desejada como parte do início de uma atualização da instância.
- Você só pode reverter para uma versão anterior de um modelo de execução se a versão for uma versão numerada específica. A opção de reversão não estará disponível se o grupo do Auto Scaling estiver configurado para usar a versão `$Latest` ou a versão do modelo de execução `$Default`.
- Você também não pode reverter para um modelo de execução configurado para usar um alias de AMI do AWS Systems Manager Parameter Store.
- A configuração salva pela última vez no grupo do Auto Scaling deve estar em um estado estável. Se não estiver num estado estável, o fluxo de trabalho de reversão ainda ocorrerá, mas eventualmente falhará. Até você resolver o problema, o grupo do Auto Scaling poderá estar em um estado de falha e não conseguir mais iniciar instâncias com êxito. Isso pode afetar a disponibilidade do serviço ou da aplicação.

Iniciar uma reversão manualmente

Console

Para iniciar manualmente a reversão de uma atualização de instância (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

3. Na guia Atualização de instância, em Atualização de instância ativa, escolha Ações, Iniciar reversão.
4. Quando a confirmação for solicitada, escolha Confirmar.

AWS CLI

Para iniciar manualmente a reversão de uma atualização de instância (AWS CLI)

Use o rollback-instance-refreshcomando do AWS CLI e forneça o nome do grupo Auto Scaling.

```
aws autoscaling rollback-instance-refresh --auto-scaling-group-name my-asg
```

Resultado do exemplo:

```
{
    "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

> **Tip**
>
> Se esse comando gerar um erro, verifique se você atualizou o AWS CLI localmente para a versão mais recente.

Iniciar uma atualização de instância com reversão automática

Usando o recurso de reversão automática, você pode reverter automaticamente a atualização da instância quando ela falhar, como quando há erros ou quando um CloudWatch alarme específico da Amazon entra no estado. `ALARM`

Se você ativar a reversão automática e houver erros ao substituir as instâncias, a atualização da instância tentará concluir todas as substituições por uma hora antes de falhar e reverter. Esses erros geralmente são causados por coisas como falhas de inicialização do EC2, verificações de integridade mal configuradas ou por não ignorar ou permitir o encerramento de instâncias que estão no estado `Standby` ou protegidas contra redução.

A especificação de CloudWatch alarmes é opcional. Para especificar um alarme, primeiro você precisa criá-lo. Você pode especificar alarmes métricos e alarmes compostos. Para obter informações sobre como criar o alarme, consulte o Guia CloudWatch do usuário da Amazon. Usando

as métricas do Elastic Load Balancing como exemplo, se você usar um Application Load Balancer, poderá usar as métricas `HTTPCode_ELB_5XX_Count` e`HTTPCode_ELB_4XX_Count`.

Considerações

- Se você especificar um CloudWatch alarme, mas não ativar a reversão automática, e o estado do alarme continuar`ALARM`, a atualização da instância falhará sem reverter.
- Você pode escolher no máximo 10 alarmes ao iniciar uma atualização da instância.
- Ao escolher um CloudWatch alarme, o alarme deve estar em um estado compatível. Se o estado do alarme for `INSUFFICIENT_DATA` ou`ALARM`, você receberá um erro ao tentar iniciar a atualização da instância.
- Ao criar um alarme para uso do Amazon EC2 Auto Scaling, o alarme deve incluir como tratar pontos de dados ausentes. Se uma métrica tiver frequentemente pontos de dados ausentes por projeto, o estado do alarme será `INSUFFICIENT_DATA` durante esses períodos. Quando isso acontece, o Amazon EC2 Auto Scaling não pode substituir instâncias até que novos pontos de dados sejam encontrados. Para forçar o alarme a manter o estado `ALARM` ou `OK` anterior, você pode optar por ignorar os dados ausentes. Para obter mais informações, consulte Configurando como os alarmes tratam dados perdidos no Guia CloudWatch do usuário da Amazon.

Console

Para iniciar uma atualização de instância com reversão automática (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.
3. Na guia Atualização de instância, em Atualização ativa de instância, escolha Iniciar atualização de instância.
4. Siga o Iniciar uma atualização de instância (console) procedimento e defina as configurações de atualização da instância conforme necessário.
5. (Opcional) Em Atualizar configurações, para CloudWatch alarme, escolha Ativar CloudWatch alarmes e, em seguida, escolha um ou mais alarmes para identificar quaisquer problemas e falhar na operação se um alarme entrar nesse estado. `ALARM`
6. Em Configurações de reversão, escolha Habilitar a reversão automática para reverter automaticamente uma atualização de instância com falha para a configuração que você salvou pela última vez no grupo do Auto Scaling antes de iniciar a atualização de instância.

7. Revise suas seleções e escolha Iniciar atualização da instância.

AWS CLI

Para iniciar uma atualização de instância com reversão automática (AWS CLI)

Use o start-instance-refreshcomando e especifique `true` a `AutoRollback` opção no`Preferences`.

O exemplo a seguir mostra como iniciar uma atualização de instância que será revertida automaticamente se ocorrer uma falha. Substitua os valores dos *`italicized`* parâmetros pelos seus próprios.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredConfiguration": {
      "LaunchTemplate": {
          "LaunchTemplateName": "my-launch-template",
          "Version": "1"
       }
    },
    "Preferences": {
      "AutoRollback": true
    }
}
```

Como alternativa, para reverter automaticamente quando a atualização da instância falhar ou quando um CloudWatch alarme especificado estiver no `ALARM` estado, especifique a `AlarmSpecification` opção no `Preferences` e forneça o nome do alarme, como no exemplo a seguir. Substitua os valores dos *`italicized`* parâmetros pelos seus próprios.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredConfiguration": {
      "LaunchTemplate": {
          "LaunchTemplateName": "my-launch-template",
```

```
            "Version": "1"
        }
    },
    "Preferences": {
      "AutoRollback": true,
      "AlarmSpecification": { "Alarms": [ "my-alarm" ] }
    }
}
```

Se houver êxito, o comando gerará uma saída semelhante à seguinte.

```
{
  "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

> **Tip**
>
> Se esse comando gerar um erro, verifique se você atualizou o AWS CLI localmente para a versão mais recente.

Usar uma atualização de instância com opção de ignorar correspondência

Ignorar a correspondência diz ao Amazon EC2 Auto Scaling para ignorar as instâncias que já tenham as atualizações mais recentes. Assim, você não substituirá mais instâncias do que o necessário. Isso é útil quando você deseja garantir que seu grupo do Auto Scaling usará uma versão específica de seu modelo de execução e substituirá apenas as instâncias que usam outra versão.

As seguintes considerações se aplicam à opção ignorar a correspondência:

- Se você iniciar uma atualização de instância com a opção de ignorar correspondência e uma configuração desejada, o Amazon EC2 Auto Scaling verificará se alguma instância corresponde à configuração desejada. Em seguida, ele substituirá apenas as instâncias que não correspondam à configuração desejada. Depois que a atualização de instância tem êxito, o Amazon EC2 Auto Scaling atualiza o grupo para refletir a configuração desejada.
- Se você iniciar uma atualização de instância com opção de ignorar correspondência, mas não especificar a configuração desejada, o Amazon EC2 Auto Scaling verificará se alguma instância corresponde à configuração que você salvou pela última vez no grupo do Auto Scaling. Em seguida, ele substituirá apenas as instâncias que não correspondam à última configuração salva.

- Você pode usar a opção de ignorar correspondência com um novo modelo de execução, uma nova versão do modelo de execução ou um conjunto de tipos de instância. Se você habilitar ignorar a correspondência, mas nenhum deles for alterado, a atualização da instância será bem-sucedida imediatamente sem substituir nenhuma instância. Se você tiver feito outras alterações na configuração desejada (como alterar a estratégia de alocação spot), o Amazon EC2 Auto Scaling aguardará a atualização de instância ser concluída com êxito. Em seguida, ele atualizará as configurações do grupo do Auto Scaling para refletir a nova configuração desejada.
- Você não pode usar ignorar a correspondência com uma nova configuração de inicialização.
- Quando você inicia uma atualização de instância e fornece uma configuração desejada, o Amazon EC2 Auto Scaling garante que todas as instâncias usem essa configuração. Portanto, quando você especifica `$Default` ou `$Latest` como a versão desejada para o seu modelo de execução e, em seguida, cria uma nova versão do modelo de execução enquanto uma atualização de instância está em andamento, todas as instâncias que já foram substituídas serão substituídas novamente.
- A ignorância de correspondência não sabe se um script de dados do usuário no modelo de execução extrairá o código atualizado e o instalará em novas instâncias. Como resultado, a ignorância de correspondência pode ignorar a substituição de instâncias que tenham códigos desatualizados instalados. Nesse caso, é necessário desativar a ignorância de correspondência para garantir que todas as instâncias recebam o código mais recente, mesmo sem uma atualização da versão do modelo de execução.

Esta seção inclui AWS CLI instruções para iniciar uma atualização de instância com a opção skip matching ativada. Para obter instruções sobre como usar o console, consulte Iniciar uma atualização de instância (console).

Ignorar correspondência (procedimento básico)

Siga as etapas desta seção para usar o AWS CLI para fazer o seguinte:

- Crie o modelo de execução que deseja aplicar às instâncias.
- Inicie uma atualização de instância para aplicar seu modelo de execução ao grupo do Auto Scaling. Se você não habilitar a opção de ignorar correspondência, todas as instâncias serão substituídas. Isso ocorre mesmo que o modelo de execução usado para provisionar a instância seja o mesmo que você especificou para a configuração desejada.

Para usar a opção de ignorar correspondência com um novo modelo de execução

1. Use o create-launch-templatecomando para criar um novo modelo de lançamento para seu grupo de Auto Scaling. Inclua a opção `--launch-template-data` e a entrada JSON que definem os detalhes das instâncias criadas para seu grupo do Auto Scaling.

   Por exemplo, use o comando a seguir para criar um modelo de execução básico com o ID de AMI *ami-0123456789abcdef0* e o tipo de instância *t2.micro*.

   ```
   aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling
    --version-description version1 \
     --launch-template-data
    '{"ImageId":"ami-0123456789abcdef0","InstanceType":"t2.micro"}'
   ```

   Se houver êxito, o comando gerará uma saída semelhante à seguinte.

   ```
   {
      "LaunchTemplate": {
        "LaunchTemplateId": "lt-068f72b729example",
        "LaunchTemplateName": "my-template-for-auto-scaling",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "CreateTime": "2023-01-30T18:16:06.000Z",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
     }
   }
   ```

   Para obter mais informações, consulte Exemplos para criar e gerenciar modelos de lançamento com o AWS CLI.

2. Use o start-instance-refreshcomando para iniciar o fluxo de trabalho de substituição da instância e aplicar seu novo modelo de lançamento com o ID*lt-068f72b729example*. Por ser novo, o modelo de execução tem apenas uma versão. Isso significa que a versão 1 do modelo de execução é o destino dessa atualização de instâncias. Se ocorrer um evento de aumento da escala na horizontal durante a atualização de instâncias e o Amazon EC2 Auto Scaling provisionar novas instâncias usando a versão 1 desse modelo de execução, elas não serão substituídas. Quando a operação for concluída com êxito, o novo modelo de execução será aplicado com êxito ao grupo do Auto Scaling.

   ```
   aws autoscaling start-instance-refresh --cli-input-json file://config.json
   ```

Conteúdo de `config.json`.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredConfiguration": {
      "LaunchTemplate": {
          "LaunchTemplateId": "lt-068f72b729example",
          "Version": "$Default"
       }
    },
    "Preferences": {
      "SkipMatching": true
    }
}
```

Se houver êxito, o comando gerará uma saída semelhante à seguinte.

```
{
  "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

Ignorar correspondências (grupos de instâncias mistas)

Se você tiver um grupo de Auto Scaling com uma política de instâncias mistas, siga as etapas nesta seção para usar o AWS CLI para iniciar uma atualização de instância com skip matching. Você tem as seguintes opções:

- Forneça um novo modelo de execução para aplicar a todos os tipos de instância especificados na política.
- Forneça um conjunto atualizado de tipos de instância alterando ou não o modelo de execução na política. Por exemplo, digamos que você faça uma migração de tipos de instância indesejados. Você usaria o modelo de execução como está, sem alterar a AMI, os grupos de segurança ou outras especificidades das instâncias a serem substituídas.

Siga as etapas em uma das seções a seguir, de acordo com a opção que atenda às suas necessidades.

Para usar a opção de ignorar correspondência com um novo modelo de execução

1. Use o create-launch-templatecomando para criar um novo modelo de lançamento para seu grupo de Auto Scaling. Inclua a opção `--launch-template-data` e a entrada JSON que definem os detalhes das instâncias criadas para seu grupo do Auto Scaling.

   Por exemplo, use o comando a seguir para criar um modelo de execução com o ID de AMI *ami-0123456789abcdef0*.

   ```
   aws ec2 create-launch-template --launch-template-name my-new-template --version-
   description version1 \
     --launch-template-data '{"ImageId":"ami-0123456789abcdef0"}'
   ```

   Se houver êxito, o comando gerará uma saída semelhante à seguinte.

   ```
   {
      "LaunchTemplate": {
        "LaunchTemplateId": "lt-04d5cc9b88example",
        "LaunchTemplateName": "my-new-template",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "CreateTime": "2023-01-31T15:56:02.000Z",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
     }
   }
   ```

   Para obter mais informações, consulte Exemplos para criar e gerenciar modelos de lançamento com o AWS CLI.

2. Para ver a política de instâncias mistas existente para seu grupo de Auto Scaling, execute o describe-auto-scaling-groupscomando. Você precisará dessas informações na próxima etapa, ao iniciar a atualização de instância.

   O comando de exemplo a seguir retorna a política de instâncias mistas configurada para o grupo do Auto Scaling chamado *my-asg*.

   ```
   aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
   ```

   Se houver êxito, o comando gerará uma saída semelhante à seguinte.

```
{
  "AutoScalingGroups":[
    {
      "AutoScalingGroupName":"my-asg",
      "AutoScalingGroupARN":"arn",
      "MixedInstancesPolicy":{
        "LaunchTemplate":{
          "LaunchTemplateSpecification":{
            "LaunchTemplateId":"lt-073693ed27example",
            "LaunchTemplateName":"my-old-template",
            "Version":"$Default"
          },
          "Overrides":[
            {
              "InstanceType":"c5.large"
            },
            {
              "InstanceType":"c5a.large"
            },
            {
              "InstanceType":"m5.large"
            },
            {
              "InstanceType":"m5a.large"
            }
          ]
        },
        "InstancesDistribution":{
          "OnDemandAllocationStrategy":"prioritized",
          "OnDemandBaseCapacity":1,
          "OnDemandPercentageAboveBaseCapacity":50,
          "SpotAllocationStrategy":"price-capacity-optimized"
        }
      },
      "MinSize":1,
      "MaxSize":5,
      "DesiredCapacity":4,
      ...
    }
  ]
}
```

3. Use o start-instance-refreshcomando para iniciar o fluxo de trabalho de substituição da instância e aplicar seu novo modelo de lançamento com o ID*lt-04d5cc9b88example*. Por ser novo, o modelo de execução tem apenas uma versão. Isso significa que a versão 1 do modelo de execução é o destino dessa atualização de instâncias. Se ocorrer um evento de aumento da escala na horizontal durante a atualização de instâncias e o Amazon EC2 Auto Scaling provisionar novas instâncias usando a versão 1 desse modelo de execução, elas não serão substituídas. Quando a operação for concluída com êxito, a política de instâncias mistas atualizada será aplicada com êxito ao grupo do Auto Scaling.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`.

```
{
  "AutoScalingGroupName":"my-asg",
  "DesiredConfiguration":{
    "MixedInstancesPolicy":{
      "LaunchTemplate":{
        "LaunchTemplateSpecification":{
          "LaunchTemplateId":"lt-04d5cc9b88example",
          "Version":"$Default"
        },
        "Overrides":[
          {
            "InstanceType":"c5.large"
          },
          {
            "InstanceType":"c5a.large"
          },
          {
            "InstanceType":"m5.large"
          },
          {
            "InstanceType":"m5a.large"
          }
        ]
      },
      "InstancesDistribution":{
        "OnDemandAllocationStrategy":"prioritized",
        "OnDemandBaseCapacity":1,
        "OnDemandPercentageAboveBaseCapacity":50,
```

```
        "SpotAllocationStrategy":"price-capacity-optimized"
        }
      }
    }
  },
  "Preferences":{
    "SkipMatching":true
  }
}
```

Se houver êxito, o comando gerará uma saída semelhante à seguinte.

```
{
  "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"
}
```

No próximo procedimento, você fornecerá um conjunto atualizado de tipos de instância sem alterar o modelo de execução.

Para usar a opção de ignorar correspondência com um conjunto atualizado de tipos de instância

1. Para ver a política de instâncias mistas existente para seu grupo de Auto Scaling, execute o describe-auto-scaling-groupscomando. Você precisará dessas informações na próxima etapa, ao iniciar a atualização de instância.

   O comando de exemplo a seguir retorna a política de instâncias mistas configurada para o grupo do Auto Scaling chamado *my-asg*.

   ```
   aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
   ```

   Se houver êxito, o comando gerará uma saída semelhante à seguinte.

   ```
   {
     "AutoScalingGroups":[
       {
         "AutoScalingGroupName":"my-asg",
         "AutoScalingGroupARN":"arn",
         "MixedInstancesPolicy":{
           "LaunchTemplate":{
             "LaunchTemplateSpecification":{
   ```

```
            "LaunchTemplateId":"lt-073693ed27example",
            "LaunchTemplateName":"my-template-for-auto-scaling",
            "Version":"$Default"
          },
          "Overrides":[
            {
              "InstanceType":"c5.large"
            },
            {
              "InstanceType":"c5a.large"
            },
            {
              "InstanceType":"m5.large"
            },
            {
              "InstanceType":"m5a.large"
            }
          ]
        },
        "InstancesDistribution":{
          "OnDemandAllocationStrategy":"prioritized",
          "OnDemandBaseCapacity":1,
          "OnDemandPercentageAboveBaseCapacity":50,
          "SpotAllocationStrategy":"price-capacity-optimized"
        }
      },
      "MinSize":1,
      "MaxSize":5,
      "DesiredCapacity":4,
      ...
    }
  ]
}
```

2. Use o start-instance-refreshcomando para iniciar o fluxo de trabalho de substituição de instâncias e aplicar suas atualizações. Para substituir instâncias que usam tipos de instância específicos, a configuração desejada deve especificar a política de instâncias mistas somente com os tipos de instância que você deseja. Você pode escolher se deseja adicionar novos tipos de instância no lugar deles.

   O comando de exemplo a seguir inicia uma atualização de instância sem o tipo de instância indesejado *`m5a.large`*. Quando um tipo de instância de seu grupo não corresponde a um dos três tipos de instância restantes, as instâncias são substituídas. (Uma atualização de

instância não escolhe os tipos de instância dos quais provisionar as novas instâncias; são as estratégias de alocação que fazem isso.) Quando a operação for concluída com êxito, a política de instâncias mistas atualizada será aplicada com êxito ao grupo do Auto Scaling.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`

```
{
  "AutoScalingGroupName":"my-asg",
  "DesiredConfiguration":{
    "MixedInstancesPolicy":{
      "LaunchTemplate":{
        "LaunchTemplateSpecification":{
          "LaunchTemplateId":"lt-073693ed27example",
          "Version":"$Default"
        },
        "Overrides":[
          {
            "InstanceType":"c5.large"
          },
          {
            "InstanceType":"c5a.large"
          },
          {
            "InstanceType":"m5.large"
          }
        ]
      },
      "InstancesDistribution":{
        "OnDemandAllocationStrategy":"prioritized",
        "OnDemandBaseCapacity":1,
        "OnDemandPercentageAboveBaseCapacity":50,
        "SpotAllocationStrategy":"price-capacity-optimized"
        }
      }
    }
  },
  "Preferences":{
    "SkipMatching":true
  }
}
```

Adicionar pontos de verificação a uma atualização de instância

Ao usar uma atualização de instância, você pode escolher substituir instâncias em fases para poder executar verificações em suas instâncias durante o uso. Para fazer uma substituição em fases, adicione pontos de verificação, que são pontos no tempo em que a atualização da instância pausa. O uso de pontos de verificação dá a você maior controle sobre como escolhe atualizar seu grupo do Auto Scaling. Isso ajuda a confirmar que sua aplicação funcionará de forma confiável e previsível.

Conteúdo

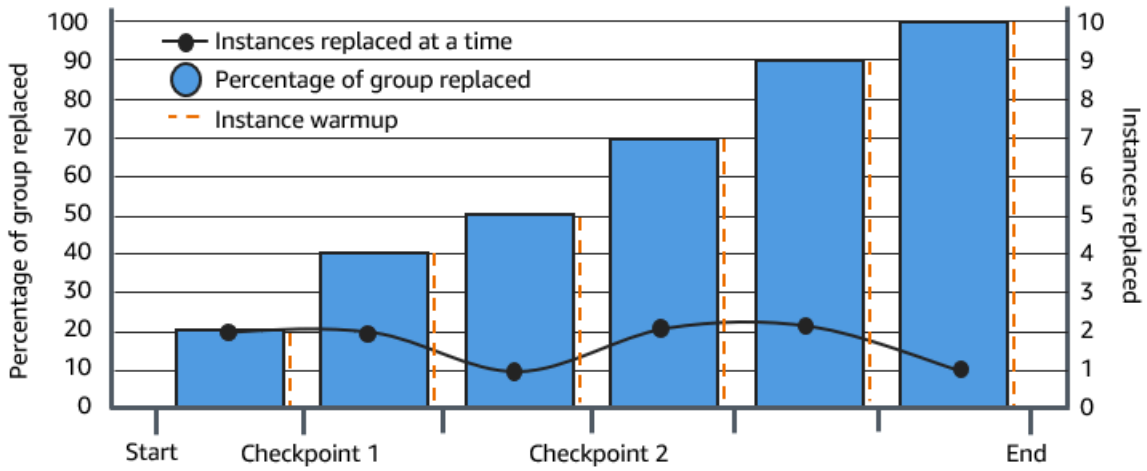
- Como funcionam
- Considerações
- Ative os pontos de verificação usando o ou Console de gerenciamento da AWS AWS CLI

Como funcionam

Ao iniciar uma atualização de instância, você especifica os pontos de controle como porcentagens do número total de instâncias no grupo do Auto Scaling. Esses pontos de controle indicam a porcentagem mínima de instâncias no grupo do Auto Scaling que devem ser novas instâncias antes que o ponto de controle seja considerado atingido. Por exemplo, se seus pontos de controle forem `[20, 50, 100]`, o primeiro ponto de controle é atingido quando 20% das instâncias são novas, o segundo quando 50% são novas e o ponto de controle final quando todas as instâncias são novas.

O Amazon EC2 Auto Scaling acelera as substituições de instâncias para respeitar as porcentagens de ponto de controle especificadas, mantendo a porcentagem mínima de integridade do grupo. Para atingir uma porcentagem de pontos de verificação, o Amazon EC2 Auto Scaling às vezes substitui menos, mas nunca mais do que a porcentagem mínima de integridade permitida.

Considere o seguinte grupo do Auto Scaling que tem 10 instâncias. As porcentagens do ponto de verificação são `[20,50,100]`, a porcentagem mínima de integridade é 80% e a porcentagem máxima de integridade é 100%. Para manter a porcentagem mínima de integridade, apenas duas instâncias podem ser substituídas por vez. O diagrama a seguir resume o processo de substituição de instâncias antes que um ponto de verificação seja alcançado.



No exemplo acima, há um período de aquecimento da instância para cada nova instância iniciada. Você também pode ter um gancho do ciclo de vida que coloca uma instância em um estado de espera e, em seguida, executa uma ação personalizada ao iniciar ou encerrar.

O Amazon EC2 Auto Scaling emite eventos para cada ponto de controle, exceto para o ponto de controle 100% completo. Você pode adicionar uma EventBridge regra para enviar os eventos para um destino, como o Amazon SNS. Assim, você é notificado quando pode executar as verificações necessárias. Para obter mais informações, consulte Crie EventBridge regras, por exemplo, eventos de atualização.

Considerações

Mantenha as seguintes considerações em mente ao usar pontos de verificação:

- Como os pontos de verificação são baseados em percentuais, o número de instâncias a serem substituídas muda de acordo com o tamanho do grupo. Quando uma atividade de aumento de escala na horizontal ocorre e o tamanho do grupo aumenta, uma operação em andamento pode chegar a um ponto de verificação novamente. Se isso acontecer, o Amazon EC2 Auto Scaling enviará outra notificação e repetirá o tempo de espera entre pontos de verificação antes de continuar.
- É possível pular um ponto de verificação sob certas circunstâncias. Por exemplo, suponha que seu grupo do Auto Scaling tenha duas instâncias e seus percentuais de ponto de verificação sejam `[10,40,100]`. Após a primeira instância ser substituída, o Amazon EC2 Auto Scaling calcula que 50% do grupo foi substituído. Como 50% é maior do que os dois primeiros pontos de verificação, ele ignora o primeiro ponto de verificação (`10`) e envia uma notificação para o segundo ponto de verificação (`40`).

- O cancelamento da operação impede que quaisquer outras substituições sejam feitas. Se a operação for cancelada ou ela falhar antes de atingir o último ponto de verificação, quaisquer instâncias que já tiverem sido substituídas não serão revertidas para a configuração anterior.
- No caso de uma atualização parcial, quando você executa novamente a operação, o Amazon EC2 Auto Scaling não é reiniciado desde o último ponto de verificação, nem para quando apenas as instâncias mais antigas são substituídas. No entanto, ele mira as instâncias mais antigas para substituição primeiro antes de lidar com as instâncias novas.
- Em alguns casos, a porcentagem real concluída pode ser maior do que a porcentagem desse ponto de verificação quando a porcentagem do ponto de verificação é muito baixa em relação ao número de instâncias no grupo. Por exemplo, suponha que a porcentagem do ponto de verificação seja de 20% e o grupo tenha quatro instâncias. Se o Amazon EC2 Auto Scaling substituir uma das quatro instâncias, a porcentagem real substituída (25%) será maior do que a porcentagem do ponto de verificação (20%).
- Depois que um ponto de verificação é atingido, o percentual total concluído não exibirá o status mais recente até que as instâncias concluam o aquecimento. Seus percentuais de ponto de verificação são `[20,50]` com um atraso no ponto de verificação de 15 minutos e um percentual mínimo de integridade de 80%. Por exemplo, suponha que seu grupo do Auto Scaling tenha 10 instâncias e faça as seguintes substituições.
  - 0:00: duas instâncias mais antigas são substituídas por novas.
  - 0:10: duas instâncias novas concluem o aquecimento.
  - 0:25: duas instâncias mais antigas são substituídas por novas. (Para manter o percentual mínimo de integridade, apenas duas instâncias são substituídas).
  - 0:35: duas instâncias novas concluem o aquecimento.
  - 0:35: uma instância mais antiga é substituída por uma nova.
  - 0:45: uma instância nova conclui o aquecimento.

  Às 0:35, a operação para de iniciar novas instâncias. O percentual concluído ainda não reflete com precisão o número de substituições concluídas (50%), porque a nova instância não terminou de aquecer. Depois que a nova instância concluir seu período de aquecimento às 0:45, o percentual completo mostrará 50%.
- Se vários pontos de verificação estiverem habilitados, o ponto de verificação de atualização da instância deverá ser definido como 100% para garantir a substituição completa de todas as instâncias. Se você definir o ponto de verificação final para um percentual menor que 100% (por exemplo, 50%), a atualização da instância deixará de substituir as instâncias após atingir esse percentual do ponto de verificação e não continuará automaticamente até 100% de conclusão.

Ative os pontos de verificação usando o ou Console de gerenciamento da AWS AWS CLI

Você pode usar o Console de gerenciamento da AWS ou AWS CLI para ativar os pontos de verificação.

Habilitar pontos de verificação (console)

Você pode habilitar pontos de verificação antes de iniciar uma atualização de instância para substituir instâncias usando uma abordagem incremental ou em fases. Isso fornece tempo adicional para verificação.

Para iniciar uma atualização de instância que usa pontos de verificação

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Instance refresh (Atualização de instância), em Active instance refresh (Atualização de instância ativa), escolha Start instance refresh (Iniciar atualização de instância).
4. Na página Start instance refresh (Iniciar atualização de instância), insira os valores aplicáveis para Minimum healthy percentage (Percentual mínimo de integridade) e Instance warmup (Aquecimento da instância).
5. Marque a caixa de seleção Enable checkpoints (Habilitar pontos de verificação).

   Isso exibe uma caixa onde você pode definir o limite percentual para o primeiro ponto de verificação.
6. Em Proceed until ____ % of the group is refreshed (Prosseguir até ____% do grupo ser atualizado), insira um número (1–100). Isso define o percentual para o primeiro ponto de verificação.
7. Para adicionar outro ponto de verificação, escolha Add checkpoint (Adicionar ponto de verificação) e, em seguida, defina o percentual para o próximo ponto de verificação.
8. Para especificar quanto tempo o Amazon EC2 Auto Scaling espera após um ponto de verificação ser atingido, atualize os campos em Wait for **1 hour** between checkpoints (Aguardar X Y entre pontos de verificação). A unidade de tempo pode ser horas, minutos ou segundos.

9. Se você tiver concluído as seleções de atualização da instância, escolha Iniciar atualização da instância.

Habilitar pontos de verificação (AWS CLI)

Para iniciar uma atualização de instância com pontos de verificação habilitados usando o AWS CLI, você precisa de um arquivo de configuração que defina os seguintes parâmetros:

- `CheckpointPercentages`: especifica valores de limites para o percentual de instâncias que serão substituídas. Esses valores de limites fornecem os pontos de verificação. Quando o percentual de instâncias substituídas e aquecidas atinge um dos limites especificados, a operação aguarda por um período especificado. Você especifica o número de segundos para esperar em `CheckpointDelay`. Quando o período de tempo especificado tiver passado, a atualização da instância continuará até atingir o próximo ponto de verificação (se aplicável).
- `CheckpointDelay`: especifica a quantidade de tempo, em segundos, para aguardar após um ponto de verificação ser atingido antes de continuar. Escolha um período que forneça tempo suficiente para executar suas verificações.

O último valor exibido na matriz `CheckpointPercentages` descreve o percentual do grupo do Auto Scaling que precisa ser substituído com êxito. A operação fará a transição para `Successful` após essa porcentagem ser substituída com êxito e cada instância ter concluído a inicialização.

Para criar vários pontos de verificação

Para criar vários pontos de verificação, use o start-instance-refreshcomando de exemplo a seguir. Este exemplo configura uma atualização de instância que atualiza inicialmente 1% do grupo do Auto Scaling. Depois de esperar 10 minutos, ele atualiza os próximos 19% e aguarda mais 10 minutos. Finalmente, ele atualiza o resto do grupo antes de concluir a operação.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 80,
```

```
      "CheckpointPercentages": [1,20,100],
      "CheckpointDelay": 600
    }
}
```

Para criar um único ponto de verificação

Para criar um único ponto de verificação, use o start-instance-refreshcomando de exemplo a seguir. Este exemplo configura uma atualização de instância que atualiza inicialmente 20% do grupo do Auto Scaling. Depois de aguardar 10 minutos, ele atualiza então o resto do grupo antes de concluir a operação.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 80,
      "CheckpointPercentages": [20,100],
      "CheckpointDelay": 600
    }
}
```

Para atualizar parcialmente o grupo do Auto Scaling

Para substituir somente uma parte do seu grupo de Auto Scaling e depois parar completamente, use o comando de exemplo start-instance-refresha seguir. Este exemplo configura uma atualização de instância que atualiza inicialmente 1% do grupo do Auto Scaling. Depois de aguardar 10 minutos, ele atualiza então os próximos 19% antes de concluir a operação.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
```

```
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 80,
      "CheckpointPercentages": [1,20],
      "CheckpointDelay": 600
    }
}
```

Substituir instâncias do Auto Scaling com base na vida útil máxima da instância

O tempo de vida máximo da instância especifica o tempo máximo (em segundos) que uma instância pode estar em serviço antes de ser terminada e substituída. Um caso de uso comum pode ser um requisito para substituir as instâncias em uma programação devido a políticas de segurança internas ou a controles de conformidade externos.

É necessário especificar um valor de pelo menos 86.400 segundos (1 dia). Para limpar um valor definido anteriormente, especifique um novo valor de 0. Essa configuração se aplica a todas as instâncias atuais e futuras do grupo do Auto Scaling.

Conteúdo

- Considerações
- Definir o tempo de vida máximo da instância
- Limitações

Considerações

Confira as seguintes considerações sobre o uso do recurso:

- Sempre que uma instância mais antiga é substituída e uma nova instância é iniciada, a nova instância usa o modelo de execução ou a configuração de execução atualmente associada ao grupo do Auto Scaling. Se o modelo de execução ou a configuração de execução especificar o ID da imagem de máquina da Amazon (AMI) de uma versão diferente da aplicação, essa versão da aplicação será implantada automaticamente.
- Definir um valor muito baixo pode fazer com que as instâncias sejam substituídas mais rapidamente do que o desejado. Em geral, o Amazon EC2 Auto Scaling substitui uma instância de cada vez, com uma pausa entre as substituições. No entanto, se o tempo de vida máximo da instância especificado não fornecer tempo suficiente para substituir cada instância individualmente,

o Amazon EC2 Auto Scaling deverá substituir mais de uma instância por vez. Várias instâncias podem ser substituídas de uma só vez, em até 10% da capacidade atual do grupo do Auto Scaling. Para evitar a substituição de muitas instâncias de uma só vez, defina um tempo de vida máximo mais longo para a instância ou use a proteção de redução da escala horizontalmente da instância para impedir temporariamente a terminação de instâncias individuais. Para obter mais informações, consulte Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância.

- Por padrão, o Amazon EC2 Auto Scaling cria uma nova atividade de escalabilidade para encerrar a instância e, em seguida, finaliza-a. Enquanto a instância estiver sendo encerrada, outra atividade de escalonamento iniciará uma nova instância. Você pode alterar esse comportamento para iniciar antes de encerrar usando uma política de manutenção de instância. Para obter mais informações, consulte Políticas de manutenção de instância.

Definir o tempo de vida máximo da instância

Quando você cria um grupo do Auto Scaling no console, não é possível configurar o tempo de vida máximo da instância. No entanto, depois que o grupo for criado, você poderá editá-lo para definir o tempo de vida máximo da instância.

Para definir o tempo de vida máximo da instância para um grupo (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling), mostrando informações sobre o grupo selecionado.
3. Na guia Detalhes, escolha Configurações avançadas, Editar.
4. Para o Maximum instance lifetime (Tempo de vida máximo da instância), insira o número máximo de segundos que uma instância pode estar em serviço.
5. Selecione Atualizar.

Na guia Activity (Atividade), em Activity history (Histórico de atividades), é possível ver a substituição de instâncias do grupo ao longo de todo seu histórico.

Para definir o tempo de vida máximo da instância para um grupo (AWS CLI)

Você também pode usar o AWS CLI para definir a vida útil máxima da instância para grupos de Auto Scaling novos ou existentes.

Para novos grupos de Auto Scaling, use o create-auto-scaling-groupcomando.

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Veja a seguir um arquivo `config.json` de exemplo que mostra um tempo de vida máximo da instância de `2592000` segundos (30 dias).

```
{
    "AutoScalingGroupName": "my-asg",
    "LaunchTemplate": {
        "LaunchTemplateName": "my-launch-template",
        "Version": "$Default"
    },
    "MinSize": 1,
    "MaxSize": 5,
    "MaxInstanceLifetime": 2592000,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
    "Tags": []
}
```

Para grupos de Auto Scaling existentes, use o update-auto-scaling-groupcomando.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-existing-asg --
max-instance-lifetime 2592000
```

Para verificar o tempo de vida máximo da instância para um grupo do Auto Scaling

Use o comando describe-auto-scaling-groups.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Limitações

- Não há garantia de que tempo de vida máximo será exato para cada instância: não há garantia de que as instâncias serão substituídas apenas no final de sua duração máxima. Em algumas situações, talvez o Amazon EC2 Auto Scaling precise iniciar a substituição de instâncias logo

após você atualizar o parâmetro de tempo de vida máximo da instância. A razão para esse comportamento é evitar a substituição de todas as instâncias ao mesmo tempo.

- Proteção de redução da escala horizontalmente da instância autorizada: O Amazon EC2 Auto Scaling fornece proteção de redução da escala horizontalmente de instância para ajudar a controlar quais instâncias podem ser terminadas. Quando essa proteção estiver ativada em uma instância, o Amazon EC2 Auto Scaling não encerrará a instância mesmo que ela tenha atingido o tempo máximo de vida útil.
- Instâncias encerradas antes da execução: quando há apenas uma instância no grupo do Auto Scaling, o recurso de vida útil máxima da instância pode resultar em uma interrupção porque o Amazon EC2 Auto Scaling encerra uma instância e, em seguida, inicia uma nova instância por padrão. Para alterar esse comportamento de iniciar antes de encerrar, consulte Políticas de manutenção de instância.

Aumente ou diminua a capacidade de computação da sua aplicação com escalonamento.

Escalabilidade é a capacidade de aumentar ou diminuir a capacidade computacional da aplicação. A escalabilidade começa com um evento ou ação de escalabilidade que instrui um grupo do Auto Scaling a iniciar ou terminar instâncias do Amazon EC2.

O Amazon EC2 Auto Scaling fornece várias maneiras para ajustar a escalabilidade para melhor atender às necessidades de suas aplicações. Como resultado, é importante que você tenha um bom entendimento da sua aplicação. Lembre-se das seguintes considerações:

- Qual perfil o Amazon EC2 Auto Scaling deve desempenhar na arquitetura da sua aplicação? É comum pensar que a escalabilidade automática seja principalmente uma maneira de aumentar e diminuir a capacidade, mas ela também é útil para manter um número estável de servidores.
- Quais restrições de custos são importantes para você? Como o Amazon EC2 Auto Scaling usa instâncias do EC2, você apenas paga pelos recursos usados. Saber suas restrições de custo ajuda você a decidir quando escalar suas aplicações e por quanto.
- Quais métricas são importantes para sua aplicação? A Amazon CloudWatch oferece suporte a várias métricas diferentes que você pode usar com seu grupo de Auto Scaling.

Conteúdo

- Escolha seu método de escalabilidade
- Definir limites de escalabilidade para seu grupo do Auto Scaling
- Definir o aquecimento padrão da instância para um grupo do Auto Scaling
- Escalabilidade manual para o Amazon EC2 Auto Scaling
- Escalabilidade programada para o Amazon EC2 Auto Scaling
- Escalabilidade dinâmica para o Amazon EC2 Auto Scaling
- Escala preditiva para o Amazon EC2 Auto Scaling
- Controlar quais instâncias do Auto Scaling serão terminadas durante uma redução de escala na horizontal
- Suspender e retomar os processos do Amazon EC2 Auto Scaling

Escolha seu método de escalabilidade

O Amazon EC2 Auto Scaling fornece várias formas de escalar seu grupo do Auto Scaling.

Manter um número fixo de instâncias

O padrão para um grupo do Auto Scaling é não ter nenhuma política de escalabilidade anexada ou ação agendada, o que faz com que ele mantenha um tamanho fixo. Depois de você ter criado seu grupo do Auto Scaling, o grupo começa executando instâncias suficientes para atender à sua capacidade desejada. Se não houver condições de escalabilidade associadas ao grupo, ele continuará mantendo a capacidade desejada, mesmo que uma instância não esteja mais íntegra. O Amazon EC2 Auto Scaling monitora a integridade de cada instância do grupo do Auto Scaling. Quando encontra uma instância que não está mais íntegra, ela a substitui por uma nova. Você pode ler uma descrição mais aprofundada deste processo em Verificações de integridade para instâncias em um grupo do Auto Scaling.

Dimensionar manualmente

A escalabilidade manual é a maneira mais básica para escalar seu grupo do Auto Scaling. Você pode atualizar a capacidade desejada do grupo de Auto Scaling ou terminar instâncias no grupo de Auto Scaling. Para obter mais informações, consulte Escalabilidade manual para o Amazon EC2 Auto Scaling.

Escala baseada em uma programação

A escalabilidade por agendamento significa que as ações de escalabilidade são realizadas automaticamente com base na data e hora. Isso é útil quando você sabe exatamente quando aumentar ou diminuir o número de instâncias em seu grupo, simplesmente porque essa necessidade surge em uma programação previsível. Para obter mais informações, consulte Escalabilidade programada para o Amazon EC2 Auto Scaling.

Escale dinamicamente com base na demanda

Uma maneira mais avançada de escalar seus recursos, usando a escalabilidade dinâmica, permite que você defina uma política de escalabilidade que redimensione dinamicamente o grupo do Auto Scaling para atender às alterações na demanda. Por exemplo, vamos supor que você tenha uma aplicação Web que atualmente é executada em duas instâncias e você queira que a utilização da CPU do grupo do Auto Scaling permaneça em cerca de 50% quando a carga na aplicação mudar. Esse método é útil para escalar à medida que ocorrem alterações no tráfego, quando você não sabe quando o tráfego será alterado. É possível configurar políticas de escalabilidade para responder por

você. Há vários tipos de políticas (ou uma combinação delas) que você pode usar para escalar em resposta a alterações no tráfego. Para obter mais informações, consulte Escalabilidade dinâmica para o Amazon EC2 Auto Scaling.

Escalar proativamente

Também é possível combinar a escalabilidade preditiva e a escalabilidade dinâmica (abordagens proativa e reativa, respectivamente) para escalar a capacidade do EC2 mais rapidamente. Use a escalabilidade preditiva para aumentar o número de instâncias do EC2 em seu grupo do Auto Scaling em antecipação aos padrões diários e semanais nos fluxos de tráfego. Para obter mais informações, consulte Escala preditiva para o Amazon EC2 Auto Scaling.

Definir limites de escalabilidade para seu grupo do Auto Scaling

Os limites de escalabilidade representam os tamanhos mínimo e máximo do grupo que você deseja para seu grupo do Auto Scaling. Você define limites separadamente para o tamanho mínimo e máximo.

É possível redimensionar a capacidade desejada do grupo para um número que esteja dentro do intervalo dos limites de tamanho mínimo e máximo. A capacidade desejada deve ser maior ou igual ao tamanho mínimo do grupo e menor ou igual ao tamanho máximo do grupo.

- Desired capacity (Capacidade desejada): representa a capacidade inicial do grupo do Auto Scaling no momento da criação. Um grupo do Auto Scaling tenta manter a capacidade desejada. Ele começa executando o número de instâncias especificado para a capacidade desejada e manterá esse número de instâncias desde que não haja políticas de escalabilidade ou ações programadas anexadas ao grupo do Auto Scaling.
- Minimum capacity (Capacidade mínima): representa o tamanho mínimo do grupo. Quando as políticas de escalabilidade estão definidas, um grupo não pode diminuir sua capacidade desejada abaixo do limite de capacidade mínima.
- Maximum capacity (Capacidade máxima): representa o tamanho máximo do grupo. Quando as políticas de escalabilidade estão definidas, um grupo não pode aumentar sua capacidade desejada acima do limite da capacidade máxima.

Os limites de tamanho mínimo e máximo também são aplicáveis aos seguintes cenários:

- Quando você define manualmente a escala do grupo do Auto Scaling mediante a atualização de sua capacidade desejada.

- Quando há a execução de ações agendadas que atualizam a capacidade desejada. Se uma ação agendada for executada sem especificar novos limites de tamanho mínimo e máximo para o grupo, os atuais limites de tamanho mínimo e máximo do grupo serão aplicados.

Um grupo do Auto Scaling sempre tenta manter sua capacidade desejada. Em casos nos quais uma instância seja encerrada inesperadamente (p. ex., devido a uma interrupção da instância spot, uma falha na verificação de integridade ou ação humana), o grupo iniciará automaticamente uma nova instância para manter a capacidade desejada.

Para gerenciar essas configurações no console

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, em Auto Scaling, escolha Auto Scaling Groups (Grupos de Auto Scaling).
3. Na página Auto Scaling groups (Grupos do Auto Scaling), marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

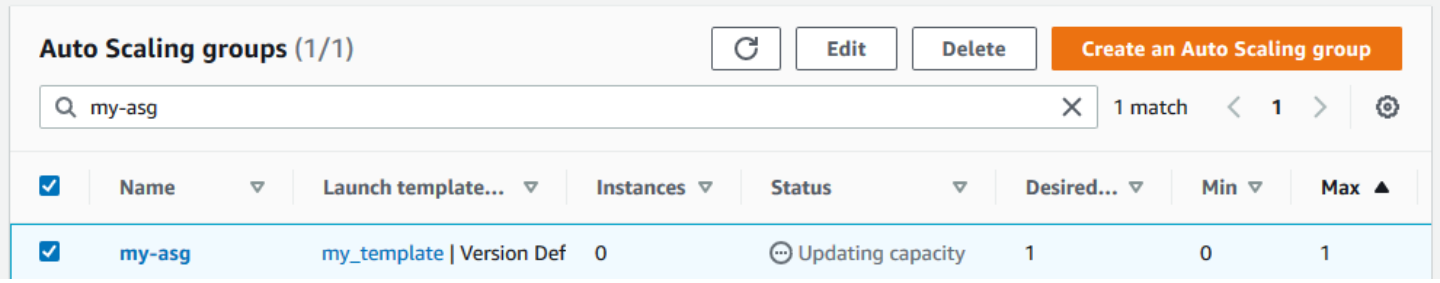
   Um painel dividido é aberto na parte inferior da página.
4. No painel inferior, na guia Detalhes, visualize ou altere as configurações atuais para as capacidades mínima, máxima e desejada do grupo. Para obter mais informações, consulte Alterar a capacidade desejada de um grupo do Auto Scaling existente.

Acima do painel Detalhes, você encontrará informações como o número atual de instâncias no grupo do Auto Scaling, as capacidades mínima, máxima e desejada, além de uma coluna de status. Se o grupo do Auto Scaling usar a ponderação da instâncias, você também encontrará o número de unidades de capacidade que contribuíram para a capacidade desejada.

Para adicionar ou remover colunas da lista, escolha o ícone de configurações na parte superior da página. Em seguida, em Auto Scaling groups attributes (Atributos dos grupos do Auto Scaling), ative ou desative cada coluna e escolha Confirm (Confirmar).

Para verificar o tamanho de seu grupo do Auto Scaling após fazer alterações.

A coluna Instances (Instâncias) exibe o número de instâncias em execução no momento. Enquanto uma instância está sendo iniciada ou terminada, a coluna Status exibe um status Updating capacity (Atualizando capacidade), conforme mostrado na imagem a seguir.



Aguarde alguns minutos e atualize a visualização para ver o status mais recente. Após a conclusão de uma atividade de escalabilidade, a coluna Instances (Instâncias) exibirá um valor atualizado.

Você pode visualizar o número de instâncias e o status das instâncas que estão em execução no momento aa guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias).

Definir o aquecimento padrão da instância para um grupo do Auto Scaling

CloudWatch coleta e agrega dados de uso, como CPU e E/S de rede, em suas instâncias do Auto Scaling. Use essas métricas para criar políticas de escalabilidade que ajustam o número de instâncias no grupo do Auto Scaling à medida que o valor da métrica selecionada aumenta e diminui.

O aquecimento padrão da instância permite que você especifique quanto tempo depois de uma instância atingir o estado `InService` ela espera antes de contribuir com dados de uso para as métricas agregadas. Esse tempo especificado é chamado de aquecimento padrão da instância. Isso impede que a escalabilidade dinâmica seja afetada por métricas para instâncias individuais que ainda não estão lidando com o tráfego de aplicações e que podem estar passando temporariamente por um uso alto de recursos de computação.

Para otimizar o desempenho de suas políticas de escalabilidade de destino e de escalabilidade de etapas, é altamente recomendável habilitar o aquecimento padrão da instância. Ele não está ativado nem configurado por padrão.

Se seu grupo do Auto Scaling estiver configurado para usar uma política de manutenção de instâncias ou se você usa uma atualização de instância para substituir instâncias, você também pode evitar que as instâncias sejam contabilizadas na porcentagem íntegra mínima antes de terminarem de inicializar quando você habilita o aquecimento padrão da instância.

Conteúdo

- Considerações sobre o desempenho de escalabilidade

- Escolha o tempo para o aquecimento de instância padrão
- Habilitar o aquecimento de instância padrão para um grupo
- Verificar o aquecimento de instância padrão para um grupo
- Encontre políticas de escalabilidade com um tempo de aquecimento da instância definido anteriormente
- Limpe o aquecimento da instância definido anteriormente para uma política de escalabilidade

Considerações sobre o desempenho de escalabilidade

A maioria das aplicações se beneficiará de ter um tempo de aquecimento de instância padrão que se aplica a todos os atributos, em vez de tempos de aquecimento diferentes para atributos diferentes. Por exemplo, se você não definir um aquecimento padrão da instância, o atributo de atualização da instância usará o período de carência da verificação de integridade como o tempo de aquecimento padrão. Se você tiver alguma política de rastreamento de destino e escalabilidade de etapas, ela usará o valor definido para o tempo de desaquecimento padrão como o tempo de aquecimento padrão. Se você tiver alguma política de escalabilidade preditiva, ela não terá um tempo de aquecimento padrão.

Enquanto as instâncias estiverem no período de aquecimento, suas políticas dinâmicas de escalabilidade só aumentarão a escala horizontalmente se o valor da métrica de instâncias que não estiver aquecendo for maior do que o limite alto de alarme da política (ou a utilização do target de uma política de escalabilidade com rastreamento do destino). Se a demanda aumentar, a intenção será reduzir a escala na horizontal de forma conservadora para proteger a disponibilidade de sua aplicação. Isto bloqueia a redução da escala horizontalmente para escalabilidade dinâmica até que as instâncias terminem de aquecer.

Ao aumentar a escala horizontalmente, o Amazon EC2 Auto Scaling considera as instâncias que estão em aquecimento como parte da capacidade do grupo ao decidir quantas instâncias adicionar ao grupo. Portanto, várias violações de alarme que exigem a adição de uma quantidade semelhante de capacidade resultam em uma única atividade de escala. A intenção é expandir de forma contínua (mas não excessivamente).

Se o aquecimento padrão da instância não estiver ativado, a quantidade de tempo que uma instância espera antes de enviar métricas CloudWatch e contá-las para a capacidade atual variará de instância para instância. Nesse caso, existe a possibilidade de suas políticas de escalabilidade funcionarem de forma imprevisível em comparação com o workload real que está ocorrendo.

Por exemplo, considere um aplicativo com um padrão de on-and-off carga de trabalho recorrente. Uma política de escalabilidade preditiva é usada para tomar decisões recorrentes sobre o aumento do número de instâncias. Como não há um tempo de aquecimento padrão para as políticas de escalabilidade preditiva, as instâncias começam a contribuir imediatamente para as métricas agregadas. Se essas instâncias tiverem maior uso de recursos no startup, a adição de instâncias poderá fazer com que as métricas agregadas tenham um pico. Dependendo do tempo necessário para o uso se estabilizar, isso pode afetar qualquer política de escalabilidade dinâmica que use essas métricas. Se o limite alto de alarme de uma política de escalabilidade dinâmica for violado, o grupo aumentará de tamanho novamente. Enquanto as novas instâncias estiverem se aquecendo, as atividades para reduzir a escala horizontalmente serão bloqueadas.

Escolha o tempo para o aquecimento de instância padrão

A chave para definir o aquecimento padrão da instância é determinar quanto tempo suas instâncias precisam terminar a inicialização e para que o consumo de recursos se estabilize após atingirem o estado`InService`. Ao escolher um tempo de aquecimento da instância, procure um equilíbrio ideal entre coletar dados de uso para tráfego legítimo, mas minimizando a coleta de dados associada a picos de uso temporário no startup.

Suponha que você tenha um grupo do Auto Scaling vinculado a um balanceador de carga do Elastic Load Balancing. Quando as instâncias concluem seu lançamento, elas são vinculadas ao balanceador de carga antes de entrarem no estado `InService`. Depois que as instâncias entram no estado `InService`, o consumo de recursos ainda pode passar por picos temporários e precisar de tempo para se estabilizar. Por exemplo, o consumo de recursos para um servidor de aplicações que precisa baixar ativos grandes e armazená-los em cache leva mais tempo para se estabilizar do que um servidor Web leve e sem ativos grandes para baixar. O aquecimento de instâncias fornece o tempo de atraso necessário para que o consumo de recursos se estabilize.

> **Important**
>
> Se você não tiver certeza de quanto tempo precisa para o aquecimento, pode começar com 300 segundos. Em seguida, diminua ou aumente gradualmente até obter o melhor desempenho de escala para sua aplicação. Talvez você precise fazer isso algumas vezes para obter os resultados desejados. Como alternativa, se você tiver alguma política de escalabilidade que tenha seu próprio tempo de aquecimento (`EstimatedInstanceWarmup`), poderá usar esse valor para começar. Para obter mais

informações, consulte Encontre políticas de escalabilidade com um tempo de aquecimento da instância definido anteriormente.

Também é necessário considerar o uso de ganchos do ciclo de vida para casos de uso em que você tem tarefas de configuração ou scripts para executar no startup. Os ganchos do ciclo de vida também podem atrasar a colocação de instâncias em serviço até que elas tenham concluído a inicialização. Eles são especialmente úteis se você tiver scripts de bootstrapping que demoram um pouco para serem concluídos. Se você adicionar um gancho do ciclo de vida, será possível reduzir o valor do aquecimento de instância padrão. Para obter mais informações sobre hooks de ciclo de vida, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

Habilitar o aquecimento de instância padrão para um grupo

É possível habilitar o aquecimento padrão da instância ao criar um grupo do Auto Scaling. Também é possível habilitar para grupos existentes.

Ao habilitar o atributo de aquecimento de instância padrão, não é mais necessário especificar valores de parâmetros de aquecimento para os seguintes recursos:

- Atualização de instância
- Escalabilidade de rastreamento de destino
- Escalabilidade em etapas

Console

Para habilitar o aquecimento de instância padrão para um novo grupo (console)

Ao criar o grupo do Auto Scaling, na página Configure advanced options (Configurar opções avançadas), em Additional settings (Configurações adicionais), selecione a opção Enable default instance warmup (Habilitar aquecimento de instância padrão). Escolha o tempo de aquecimento necessário para sua aplicação.

AWS CLI

Para habilitar o aquecimento de instância padrão para um novo grupo (AWS CLI)

Para habilitar o aquecimento de instância padrão para um grupo do Auto Scaling, adicione a opção `--default-instance-warmup` e especifique um valor, em segundos, de 0 a 3600. Depois de habilitado, um valor de `-1` desativará essa configuração.

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling com o nome *my-asg* e ativa o aquecimento padrão da instância com um valor de segundos. *120*

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg --
default-instance-warmup 120 ...
```

> **Tip**
>
> Se esse comando gerar um erro, verifique se você atualizou o AWS CLI localmente para a versão mais recente.

Console

Para habilitar o aquecimento de instância padrão para um grupo existente (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a mesma Região da AWS na qual você criou o grupo do Auto Scaling.
3. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
4. Na guia Detalhes, escolha Configurações avançadas, Editar.
5. Em Default instance warmup (Aquecimento de instância padrão), escolha o tempo de aquecimento necessário para sua aplicação.
6. Selecione Atualizar.

AWS CLI

Para habilitar o aquecimento de instância padrão para um grupo existente (AWS CLI)

O exemplo a seguir usa o update-auto-scaling-groupcomando para ativar o aquecimento padrão da instância com um valor de *120* segundos para um grupo de Auto Scaling existente chamado. *my-asg*

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --
default-instance-warmup 120
```

> **Tip**
>
> Se esse comando gerar um erro, verifique se você atualizou o AWS CLI localmente para a versão mais recente.

Verificar o aquecimento de instância padrão para um grupo

Use o procedimento a seguir para verificar o tempo de aquecimento padrão da instância para um grupo do Auto Scaling usando a função AWS CLI.

Para verificar o aquecimento padrão da instância para um grupo do Auto Scaling ()

Use o seguinte comando describe-auto-scaling-groups: *my-asg*Substitua pelo nome do seu grupo de Auto Scaling.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

O seguinte é um exemplo de resposta.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            ...
            "DefaultInstanceWarmup": 120
        }
    ]
}
```

Encontre políticas de escalabilidade com um tempo de aquecimento da instância definido anteriormente

Para identificar se você tem políticas que têm seu próprio tempo de aquecimento para,`EstimatedInstanceWarmup` execute o seguinte comando describe-policies usando o AWS CLI. *`my-asg`*Substitua pelo nome do seu grupo de Auto Scaling.

```
aws autoscaling describe-policies --auto-scaling-group-name my-asg
  --query 'ScalingPolicies[?EstimatedInstanceWarmup!=`null`]'
```

O seguinte é um exemplo de saída.

```
[
  {
    "AutoScalingGroupName":"my-asg",
    "PolicyName":"cpu50-target-tracking-scaling-policy",
    "PolicyARN":"arn",
    "PolicyType":"TargetTrackingScaling",
    "StepAdjustments":[],
    "EstimatedInstanceWarmup":120,
    "Alarms":[{
        "AlarmARN": "arn:aws:cloudwatch:us-west-2:123456789012:alarm:TargetTracking-my-
asg-AlarmHigh-fc0e4183-23ac-497e-9992-691c9980c38e",
        "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
      },
      {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
            "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
    }],
    "TargetTrackingConfiguration":{
      "PredefinedMetricSpecification":{
        "PredefinedMetricType":"ASGAverageCPUUtilization"
      },
      "TargetValue":50.0,
      "DisableScaleIn":false
    },
    "Enabled":true
  },
```

```
    ... additional policies ...

]
```

Limpe o aquecimento da instância definido anteriormente para uma política de escalabilidade

Depois de ativar o aquecimento padrão da instância, atualize todas as políticas de escalabilidade que ainda tenham seu próprio tempo de aquecimento para limpar o valor definido anteriormente. Caso contrário, ele substituirá o aquecimento padrão da instância.

Você pode atualizar as políticas de escalabilidade usando o console, AWS CLI, ou AWS SDKs. Esta seção aborda as etapas do console. Se você usar o AWS CLI ou AWS SDKs, certifique-se de preservar a configuração de política existente, mas remova a `EstimatedInstanceWarmup` propriedade. Quando você atualiza uma política de escalabilidade existente, a política será substituída pelo que você especifica ao chamar programaticamente. PutScalingPolicy Os valores originais não são mantidos.

Para limpar o aquecimento da instância definido anteriormente para uma política de escalabilidade (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Escabilidade automática, em Políticas de escalabilidade dinâmica, escolha a política na qual você está interessado e, em seguida, escolha Ações, Editar.
4. Em Aquecimento da instância, limpe o valor de aquecimento da instância para usar o valor padrão de aquecimento da instância.
5. Selecione Atualizar.

Escalabilidade manual para o Amazon EC2 Auto Scaling

Você pode ajustar manualmente o número de instâncias do EC2 em seu grupo do Auto Scaling a qualquer momento. Esse processo de alterar a contagem de instâncias manualmente é chamado

de escala manual. O ajuste de escala manual é uma alternativa ao ajuste de escala automático, especialmente se você quiser fazer alterações únicas na capacidade.

Depois de escalar manualmente o seu grupo, o Amazon EC2 Auto Scaling retoma as atividades normais de ajuste de escala automático com base nas políticas de escala e ações agendadas que você definiu. Para grupos com o aquecimento de instância padrão ativado, todas as novas instâncias passam por um período de aquecimento antes de começarem a contribuir para as métricas usadas no ajuste de escala automático. Esse período de aquecimento ajuda a estabilizar o grupo na nova capacidade. Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Às vezes, você pode querer desativar temporariamente as políticas de escala e as ações agendadas antes de escalar manualmente um grupo Isso evita o surgimento de conflitos entre as ações de escala manual e as atividades de escala automatizada. Para obter mais informações, consulte Desativar atividades de escala.

Conteúdo

- Alterar a capacidade desejada de um grupo do Auto Scaling existente
- Encerrar uma instância no seu grupo do Auto Scaling (AWS CLI)

Alterar a capacidade desejada de um grupo do Auto Scaling existente

Quando você altera a capacidade desejada de seu grupo do Auto Scaling, o Amazon EC2 Auto Scaling gerencia o processo de início ou de término de instâncias para manter o novo tamanho do grupo.

Console

Para alterar o tamanho de seu grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Detalhes, escolha Detalhes do grupo, Editar.
4. Em Desired capacity (Capacidade desejada), aumente a capacidade desejada para um. Por exemplo, para aumentar o tamanho do grupo em um, se o valor atual for 1, digite 2.

Se o novo valor para capacidade desejada for maior que a capacidade mínima desejada e a capacidade máxima desejada, a capacidade máxima desejada será automaticamente aumentada para o novo valor de capacidade desejada.

5. Quando terminar, escolha Atualizar.

Verifique se o tamanho do grupo que você especificou resultou na mesma quantidade de instâncias sendo executadas. Por exemplo, se você aumentou o tamanho do grupo em um, verifique se o grupo do Auto Scaling executou uma instância adicional.

Para verificar se o tamanho do grupo do Auto Scaling foi alterado

1. Na guia Atividade, em Histórico de atividades, você pode visualizar o progresso das atividades associadas ao grupo do Auto Scaling. A coluna Status mostra o status atual de sua instância. Enquanto sua instância está ativando, a coluna de status mostra `Not yet in service`. O status muda para `Successful` depois que a instância é ativada. Você também pode usar o botão Atualizar para ver o status atual de sua instância. Para obter mais informações, consulte Verificar uma ação de escalabilidade para um grupo do Auto Scaling.
2. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), é possível visualizar o status da instância. Demora um pouco para iniciar uma instância.
   - A guia Lifecycle (Ciclo de vida) mostra o estado de sua instância. Inicialmente, sua instância está no estado `Pending`. Quando uma instância está pronta para receber tráfego, seu estado é `InService`.
   - A coluna Health Status (Status de integridade) mostra o resultado das verificações de integridade do Amazon EC2 Auto Scaling em suas instâncias.

AWS CLI

O exemplo a seguir pressupõe que você criou um grupo do Auto Scaling com um tamanho mínimo de 1 e um tamanho máximo de 5. Portanto, o grupo atualmente tem uma instância em execução.

Para alterar o tamanho de seu grupo do Auto Scaling

Use o set-desired-capacitycomando para alterar o tamanho do seu grupo de Auto Scaling, conforme mostrado no exemplo a seguir.

```
aws autoscaling set-desired-capacity --auto-scaling-group-name my-asg \
  --desired-capacity 2
```

Se você optar por cumprir o período de desaquecimento padrão para seu grupo do Auto Scaling, especifique a opção `--honor-cooldown`, conforme mostrado no exemplo a seguir. Para obter mais informações, consulte Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling.

```
aws autoscaling set-desired-capacity --auto-scaling-group-name my-asg \
  --desired-capacity 2 --honor-cooldown
```

Para verificar o tamanho de seu grupo do Auto Scaling

Use o describe-auto-scaling-groupscomando para confirmar que o tamanho do seu grupo de Auto Scaling foi alterado, como no exemplo a seguir.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Veja a seguir o exemplo de resultado, com detalhes sobre o grupo e instâncias executadas.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "MinSize": 1,
            "MaxSize": 5,
            "DesiredCapacity": 2,
            "DefaultCooldown": 300,
            "AvailabilityZones": [
                "us-west-2a"
            ],
            "LoadBalancerNames": [],
            "TargetGroupARNs": [],
            "HealthCheckType": "EC2",
            "HealthCheckGracePeriod": 300,
            "Instances": [
```

```
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-05b4f7d5be44822a6",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "Pending"
                },
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-0c20ac468fa3049e8",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService"
                }
            ],
            "CreatedTime": "2019-03-18T23:30:42.611Z",
            "SuspendedProcesses": [],
            "VPCZoneIdentifier": "subnet-c87f2be0",
            "EnabledMetrics": [],
            "Tags": [],
            "TerminationPolicies": [
                "Default"
            ],
            "NewInstancesProtectedFromScaleIn": false,
            "ServiceLinkedRoleARN": "arn",
            "TrafficSources": []
        }
    ]
}
```

Observe que `DesiredCapacity` mostra o novo valor. Seu grupo do Auto Scaling iniciou uma instância adicional.

Encerrar uma instância no seu grupo do Auto Scaling (AWS CLI)

Há momentos em que talvez você queira manualmente reduzir a escala horizontalmente em seu grupo do Auto Scaling, mas queira e encerrar uma instância específica. Você pode escalar manualmente em seu grupo de Auto Scaling usando o comando terminate-instance-in-auto-scaling-group e especificando o ID da instância que você quer encerrar e a `--should-decrement-desired-capacity` opção, conforme mostrado no exemplo a seguir.

```
aws autoscaling terminate-instance-in-auto-scaling-group \
  --instance-id i-026e4c9f62c3e448c --should-decrement-desired-capacity
```

Este é um exemplo de saída, com detalhes sobre a atividade de escalabilidade.

```
{
    "Activities": [
        {
            "ActivityId": "b8d62b03-10d8-9df4-7377-e464ab6bd0cb",
            "AutoScalingGroupName": "my-asg",
            "Description": "Terminating EC2 instance: i-026e4c9f62c3e448c",
            "Cause": "At 2023-09-23T06:39:59Z instance i-026e4c9f62c3e448c was taken
 out of service in response to a user request, shrinking the capacity from 1 to 0.",
            "StartTime": "2023-09-23T06:39:59.015000+00:00",
            "StatusCode": "InProgress",
            "Progress": 0,
            "Details": "{\"Subnet ID\":\"subnet-6194ea3b\",\"Availability Zone\":\"us-
west-2c\"}"
        }
    ]
}
```

Esta opção não está disponível no console. No entanto, você pode usar a página Instâncias do console do Amazon EC2 para encerrar uma instância no seu grupo do Auto Scaling. Quando você faz isso, o Amazon EC2 Auto Scaling detecta que a instância não está mais em execução e a substitui automaticamente como parte do processo de verificação de integridade. Leva um ou dois minutos após a conclusão da instância para que uma nova instância seja iniciada. Consulte

mais informações sobre como terminar uma instância em Como terminar uma instância no Guia do usuário do Amazon EC2.

Observe que, se você encerrar instâncias em seu grupo causando distribuição desigual entre zonas, o Amazon EC2 Auto Scaling reequilibrará o grupo para restabelecer uma distribuição uniforme, a menos que você suspenda o processo`AZRebalance`. Para obter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.

Escalabilidade programada para o Amazon EC2 Auto Scaling

Com o Application Auto Scaling, você pode definir sua própria programação de ajuste de escala de acordo com alterações de carga previsíveis. Você cria ações programadas que aumentam ou diminuem a capacidade desejada do seu grupo em horários específicos.

Por exemplo, suponhamos que você experiencie um padrão de tráfego semanal regular, em que a carga aumenta no meio da semana e diminui no final da semana. É possível configurar uma escalabilidade programada no Amazon EC2 Auto Scaling que se alinhe a este padrão:

- Na manhã de quarta-feira, uma ação programada amplia a capacidade ao aumentar a capacidade mínima previamente definida do destino escalável.
- Na noite de sexta-feira, outra ação programada reduz a capacidade ao diminuir a capacidade máxima previamente definida do destino escalável.

Essas ações de escalabilidade programadas permitem otimizar os custos e a performance. A aplicação tem capacidade suficiente para lidar com o pico de tráfego no meio da semana, mas não faz provisionamento excessivo de capacidade desnecessária em outros momentos.

É possível usar a escalabilidade programada e as políticas de escalabilidade em conjunto para obter os benefícios de abordagens proativas e reativas para a escalabilidade. Após a execução de uma ação de escalabilidade programada, a política de escalabilidade pode continuar a tomar decisões sobre a necessidade de escalar ainda mais a capacidade. Isso ajuda a garantir que você tenha capacidade suficiente para lidar com a carga de sua aplicação. Embora sua aplicação seja escalada para atender à demanda, a capacidade atual deve estar dentro das capacidades mínima e máxima definidas pela ação agendada.

Conteúdo

- Como a escalabilidade programada funciona

- Programações recorrentes
- Fuso horário
- Considerações
- Limitações
- Criar uma ação programada
- Visualizar detalhes da ação programada
- Excluir uma ação programada

Como a escalabilidade programada funciona

Para usar a escalabilidade programada, crie ações programadas, que instruem o Amazon EC2 Auto Scaling a executar ações de escalabilidade em momentos específicos. Ao criar uma ação programada, você especifica o destino escalável, quando a ação de escalabilidade deve ocorrer, a capacidade mínima e a capacidade máxima. Você pode criar ações agendadas que escalam somente uma vez ou que escalam em uma programação recorrente.

No momento especificado, o Amazon EC2 Auto Scaling escala com base nos novos valores de capacidade, comparando a capacidade atual com a capacidade desejada especificada.

- Se a capacidade atual for menor do que a capacidade desejada especificada, o Amazon EC2 Auto Scaling aumentará a escala horizontalmente ou adicionará instâncias à capacidade desejada especificada.
- Se a capacidade atual for maior do que a capacidade desejada especificada, o Amazon EC2 Auto Scaling se reduzirá a escala horizontalmente ou removerá instâncias para a capacidade desejada especificada.

Uma ação programada define os tamanhos mínimo, máximo e desejado para o que é especificado pela ação programada na data e horário especificado. Você pode criar uma ação programada para apenas uma dessas capacidades de cada vez, por exemplo, a capacidade desejada.. Em alguns casos, no entanto, é necessário incluir os tamanhos mínimo e máximo para garantir que a nova capacidade desejada especificada na ação não esteja fora desses limites.

Programações recorrentes

Para criar uma agenda recorrente usando o AWS CLI ou um SDK, especifique uma expressão cron e um fuso horário para descrever quando essa ação agendada deve ocorrer novamente.

Opcionalmente, você pode especificar uma data e hora para a hora de início, a hora de término ou ambas.

Para criar uma agenda recorrente usando o Console de gerenciamento da AWS, especifique o padrão de recorrência, o fuso horário, a hora de início e a hora de término opcional da ação agendada. Todas as opções de padrão de recorrência são baseadas em expressões do cron. Alternativamente, você pode escrever sua própria expressão do cron personalizada.

A expressão do cron consiste em cinco campos separados por espaços: [Minuto] [Hora] [Dia_do_mês] [Mês_do_ano] [Dia_da_semana]. Por exemplo, a expressão do cron `30 6 * * 2` configura uma ação programada que se repete todas as terças-feiras às 6h30. O asterisco é usado como um curinga para corresponder a todos os valores de um campo. Para outros exemplos de expressões cron, consulte https://crontab.guru/examples.html. Para obter informações sobre como gravar suas próprias expressões do cron nesse formato, consulte Crontab.

Selecione os horários de início e término cuidadosamente. Lembre-se do seguinte:

- Se você especificar uma hora de início, o Amazon EC2 Auto Scaling executará a ação nessa hora, e depois executará a ação de acordo com a recorrência especificada.
- Se você especificar um horário de término, a ação não será mais repetida após esse horário. A ação programada não se manterá na sua conta depois que ela tiver chegado ao fim.
- Se uma hora de recorrência corresponder exatamente à hora de término, o Amazon EC2 Auto Scaling não executará a ação programada na hora de término.
- A hora de início e a hora de término devem ser definidas em UTC quando você usa o AWS CLI ou um SDK.

Fuso horário

Por padrão, as programações recorrentes definidas por você estão no fuso horário UTC (Tempo Universal Coordenado). É possível alterar o fuso para corresponder a seu fuso horário local ou a um fuso horário de outra parte da rede. Se você especificar um o fuso horário que siga o horário de verão, ele se ajustará automaticamente ao horário de verão (DST).

Os valores válidos são os nomes canônicos para fusos horários do banco de dados de fusos horários da IANA (Internet Assigned Numbers Authority). Por exemplo, o horário do leste dos EUA é canonicamente identificado como `America/New_York`. Para obter mais informações, consulte https://www.iana.org/time-zones.

Fusos horários baseados em localização, como `America/New_York`, são ajustados automaticamente para o horário de verão (DST). No entanto, um fuso horário baseado em UTC, como `Etc/UTC`, é uma hora absoluta e não se ajustará para o horário de verão.

Por exemplo, você tem uma programação recorrente cujo fuso horário é `America/New_York`. A primeira ação de escalabilidade acontece no fuso horário `America/New_York`, antes do horário de verão ser iniciado. A próxima ação de escalabilidade acontece no fuso horário `America/New_York`, depois do horário de verão ser iniciado. A primeira ação começa às 8:00 UTC-5 na hora local, enquanto a segunda vez começa às 8:00 UTC-4 no horário local.

Se você criar uma ação agendada usando o Console de gerenciamento da AWS e especificar um fuso horário que observe o horário de verão, tanto a programação recorrente quanto os horários de início e término se ajustarão automaticamente para o horário de verão.

Considerações

Ao criar uma ação programada, lembre-se do seguinte:

- A ordem de execução das ações programadas é garantida no mesmo grupo, mas não das ações programadas entre grupos.
- Uma ação programada geralmente é executada em segundos. No entanto, a ação pode ser atrasada em até dois minutos da hora de início programada. Como as ações programadas em um grupo do Auto Scaling são executadas na ordem em que são especificadas, as ações com horas de início programadas próximas umas das outras podem demorar mais para serem executadas.
- Você pode desativar temporariamente a escalabilidade programada para um grupo do Auto Scaling, suspendendo o processo `ScheduledActions`. Isso ajuda você a impedir que ações programadas fiquem ativas sem precisar excluí-las. Em seguida, você pode retomar a escalabilidade programada quando quiser usá-la novamente. Para obter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.
- Depois de criar uma ação programada, você pode atualizar qualquer uma de suas configurações, exceto o nome.
- Quando várias ações agendadas no mesmo grupo do Auto Scaling têm expressões cron idênticas, a ordem de execução se torna arbitrária e indefinida. Para garantir um comportamento previsível, é possível usar horários de início programados exclusivos para cada ação agendada.

Limitações

- Os nomes das ações programadas devem ser exclusivos por grupo do Auto Scaling.
- A ação programada deve ter um valor de tempo exclusivo. Se você tentar programar uma atividade em um momento em que outra atividade de escalabilidade já esteja programada, a chamada será rejeitada e retornará um erro, indicando que já existe uma ação programada com essa hora de início programada.
- Você pode criar um máximo de 125 ações programadas por grupo do Auto Scaling.

Criar uma ação programada

Para criar uma ação programada para o grupo do Auto Scaling, use um dos métodos a seguir:

Console

Para criar uma ação programada

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Automatic scaling (Escalabilidade automática), em Scheduled actions (Ações programadas), escolha Create scheduled action (Criar ação programada).
4. Insira um Name (Nome), para a ação programada.
5. Em Capacidade desejada, Mín., Máx., escolha a nova capacidade desejada do grupo e os novos limites de tamanho mínimo e máximo. A capacidade desejada deve ser maior ou igual ao tamanho mínimo do grupo e menor ou igual ao tamanho máximo do grupo.
6. Em Recurrence (Recorrência), selecione uma das opções disponíveis.

   - Se você quiser escalar em uma programação recorrente, escolha com que frequência o Amazon EC2 Auto Scaling deve executar a ação programada.
     - Se você escolher uma opção que começa com Every (A cada), a expressão Cron será criada para você.
     - Se você escolher Cron, insira uma expressão do cron que especifique quando executar a ação, em UTC.

- Se você quiser escalar apenas uma vez, escolha Once (Uma vez).

7. Em Time zone (Fuso horário), escolha um fuso horário. O padrão é `Etc/UTC`.

   Todos os fusos horários listados são do banco de dados de fuso horário da IANA. Para obter mais informações, consulte https://en.wikipedia. org/wiki/List_of_tz_database_time_zones.

8. Defina uma data e hora para Specific start time (Horário de início específico).

   - Se você escolher uma programação recorrente, o horário inicial definirá quando a primeira ação programada na série recorrente será executada.
   - Se você escolheu Once (Uma vez) como recorrência, o horário inicial define a data e a hora para a ação programada ser executada.

9. (Opcional) Para programações recorrentes, você pode especificar uma hora final escolhendo Set End Time (Definir horário de término) e, em seguida, escolher uma data e hora para End by (Encerrar em).

10. Escolha Criar. O console exibe as ações programadas para o grupo do Auto Scaling.

AWS CLI

Para criar uma ação programada, você pode usar um dos seguintes comandos de exemplo. Substitua cada *user input placeholder* por suas próprias informações.

Exemplo: para escalar apenas uma vez

Use o seguinte comando put-scheduled-update-group-action com as `--desired-capacity` opções `--start-time "YYYY-MM-DDThh:mm:ssZ"` e.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-one-time-action \
  --auto-scaling-group-name my-asg --start-time "2021-03-31T08:00:00Z" --desired-capacity 3
```

Para programar a escalabilidade em uma programação recorrente

Use o seguinte comando put-scheduled-update-group-action com as `--desired-capacity` opções `--recurrence "cron expression"` e.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-recurring-action \
```

```
  --auto-scaling-group-name my-asg --recurrence "0 9 * * *" --desired-capacity 3
```

Por padrão, o Amazon EC2 Auto Scaling executa o cronograma de recorrência especificado com base no fuso horário UTC. Para especificar um fuso horário diferente, inclua a opção `--time-zone` e especifique o nome canônico do fuso horário IANA, como no exemplo a seguir.

```
--time-zone "America/New_York"
```

Para obter mais informações, consulte https://en.wikipedia. org/wiki/List_of_tz_database_time_zones.

Visualizar detalhes da ação programada

Para visualizar detalhes das próximas ações programadas para o grupo do Auto Scaling, use um dos métodos a seguir:

Console

Para visualizar os detalhes da escala de plantão

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Selecione seu grupo do Auto Scaling.
3. Na guia Automatic scaling (Escalabilidade automática) em Scheduled actions (Ações programadas), selecione uma ação programada.

Observe que o console mostra os valores de Start time (Horário de início) e End time (Horário de término) para o seu horário local com o deslocamento de UTC em vigor na data e horário especificados. O deslocamento de UTC é a diferença, em horas e minutos, da hora local para a UTC. O valorde Time zone (Fuso horário) mostra seu fuso horário solicitado, por exemplo, `America/New_York`.

AWS CLI

Use o seguinte comando describe-scheduled-actions:

```
aws autoscaling describe-scheduled-actions --auto-scaling-group-name my-asg
```

Se houver êxito, o comando gerará uma saída semelhante à seguinte.

```
{
  "ScheduledUpdateGroupActions": [
    {
      "AutoScalingGroupName": "my-asg",
      "ScheduledActionName": "my-recurring-action",
      "Recurrence": "30 0 1 1,6,12 *",
      "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledUpdateGroupAction:8e86b655-b2e6-4410-8f29-
b4f094d6871c:autoScalingGroupName/my-asg:scheduledActionName/my-recurring-action",
      "StartTime": "2020-12-01T00:30:00Z",
      "Time": "2020-12-01T00:30:00Z",
      "MinSize": 1,
      "MaxSize": 6,
      "DesiredCapacity": 4
    }
  ]
}
```

Verificar as atividades de escalabilidade

Para verificar as atividades de escalabilidade associadas à escalabilidade programada, consulteVerificar uma ação de escalabilidade para um grupo do Auto Scaling.

Excluir uma ação programada

Para excluir uma ação programada, use um dos métodos a seguir:

Console

Para excluir uma ação programada

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Selecione seu grupo do Auto Scaling.
3. Na guia Automatic scaling (Escalabilidade automática) em Scheduled actions (Ações programadas), selecione uma ação programada.
4. Selecione Ações, Excluir.

5. Quando a confirmação for solicitada, escolha Sim, excluir.

AWS CLI

Use o seguinte comando delete-scheduled-action:

```
aws autoscaling delete-scheduled-action --auto-scaling-group-name my-asg \
  --scheduled-action-name my-recurring-action
```

Escalabilidade dinâmica para o Amazon EC2 Auto Scaling

A escalabilidade dinâmica dimensiona a capacidade do seu grupo do Auto Scaling de acordo com a ocorrência de alterações no tráfego.

O Amazon EC2 Auto Scaling oferece suporte aos seguintes tipos de políticas de escalabilidade dinâmica:

- Escalabilidade de rastreamento de metas — aumente e diminua a capacidade atual do grupo com base em uma CloudWatch métrica da Amazon e em um valor alvo. Esse tipo de política funciona de modo semelhante ao seu termostato em casa. Você escolhe uma temperatura e o termostato faz o resto.
- Escalabilidade em etapas: aumenta e diminui a capacidade atual do grupo com base em um conjunto de ajustes de escalabilidade, conhecidos como ajustes em etapas, que variam de acordo com o porte da violação do alarme.
- Escalabilidade simples: aumenta e diminui a capacidade atual do grupo com base em um único ajuste de escalabilidade, com um período de esfriamento entre cada atividade de escalonamento.

Recomendamos enfaticamente que você use políticas de escala de monitoramento do destino e escolha uma métrica que mude de forma inversamente proporcional a uma mudança na capacidade do seu grupo do Auto Scaling. Portanto, se você dobrar o tamanho do seu grupo do Auto Scaling, a métrica diminuirá em 50%. Isso permite que os dados de métricas acionem com precisão eventos de escalabilidade proporcionais. Estão incluídas métricas como a utilização média da CPU ou a contagem média de solicitações por destino.

Com o monitoramento de objetivos, um grupo do Auto Scaling reduz a escala horizontalmente em proporção direta à carga real em sua aplicação. Isso significa que, além de atender à necessidade

imediata de capacidade em resposta a mudanças de carga, uma política de monitoramento de objetivo também pode se adaptar às mudanças de carga que ocorram ao longo do tempo, p. ex., em decorrência de variações sazonais.

As políticas de rastreamento de metas também eliminam a necessidade de definir manualmente CloudWatch alarmes e ajustes de escala. O Amazon EC2 Auto Scaling lida com isso automaticamente com base no destino definido.

Conteúdo

- Como funcionam as políticas de escalabilidade dinâmica
- Várias políticas de escalabilidade dinâmica
- Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling
- Políticas de escalabilidade simples e em etapas do Amazon EC2 Auto Scaling
- Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling
- Ajuste de escala baseado no Amazon SQS
- Verificar uma ação de escalabilidade para um grupo do Auto Scaling
- Desabilitar uma política de escalabilidade para um grupo do Auto Scaling
- Desabilitar uma política de escalabilidade para um grupo do Auto Scaling
- Exemplo de políticas de escalabilidade para o AWS CLI

Como funcionam as políticas de escalabilidade dinâmica

Uma política de escalabilidade dinâmica instrui o Amazon EC2 Auto Scaling a rastrear uma métrica CloudWatch específica e define a ação a ser tomada quando o CloudWatch alarme associado está em ALARM. As métricas usadas para invocar um estado de alarme são uma agregação de métricas provenientes de todas as instâncias do grupo do Auto Scaling. (Por exemplo, vamos supor que você tenha um grupo do Auto Scaling com duas instâncias em que uma instância está com 60% de CPU e, a outra, com 40% de CPU. Na média, elas estão com 50% de CPU.) Quando a política está em vigor, o Amazon EC2 Auto Scaling ajusta a capacidade desejada do grupo para mais ou para menos quando o limite de um alarme é violado.

Quando uma política de escalabilidade dinâmica é invocada, se o cálculo de capacidade produzir um número fora do intervalo de tamanho mínimo e máximo do grupo, o Amazon EC2 Auto Scaling garantirá que a nova capacidade nunca saia dos limites de tamanho mínimo e máximo. A capacidade

é medida de duas formas: usando as mesmas unidades que você escolheu ao definir a capacidade desejada em termos de instâncias ou usando as unidades de capacidade (se o peso de instâncias for aplicado).

- Exemplo 1: um grupo do Auto Scaling tem uma capacidade máxima de 3, uma capacidade atual de 2 e uma política de escalabilidade dinâmica que adiciona 3 instâncias. Ao invocar essa política, o Amazon EC2 Auto Scaling adiciona apenas 1 instância ao grupo para impedir que ele exceda seu tamanho máximo.
- Exemplo 2: um grupo do Auto Scaling tem uma capacidade mínima de 2, uma capacidade atual de 3 e uma política de escalabilidade dinâmica que remove 2 instâncias. Ao invocar essa política, o Amazon EC2 Auto Scaling remove somente 1 instância do grupo para impedir que ele fique menor do que seu tamanho mínimo.

Quando a capacidade desejada atingir o limite de tamanho máximo, a expansão é interrompida. Se a demanda cair e a capacidade diminuir, o Amazon EC2 Auto Scaling poderá aumentar a escala na horizontal novamente.

A exceção é quando você usa a ponderação de instâncias. Nesse caso, o Amazon EC2 Auto Scaling pode aumentar a escala na horizontal acima do limite de tamanho máximo, mas somente até o peso máximo da instância. Sua intenção é chegar o mais próximo possível da nova capacidade desejada, mas ainda seguir as estratégias de alocação especificadas para o grupo. As estratégias de alocação determinam quais tipos de instância serão executados. Os pesos determinam quantas com unidades de capacidade cada instância contribui para a capacidade desejada do grupo com base no seu tipo de instância.

- Exemplo 3: um grupo do Auto Scaling tem uma capacidade máxima de 12, uma capacidade atual de 10 e uma política de escalabilidade dinâmica que adiciona 5 unidades de capacidade. Os tipos de instância têm um dos três pesos atribuídos: 1, 4 ou 6. Ao invocar a política, o Amazon EC2 Auto Scaling opta por iniciar um tipo de instância com um peso de 6 com base na estratégia de alocação. O resultado desse evento de expansão é um grupo com uma capacidade desejada de 12 e uma capacidade atual de 16.

Várias políticas de escalabilidade dinâmica

Na maioria dos casos, uma política de escalabilidade com monitoramento do objetivo é suficiente para configurar o grupo do Auto Scaling para aumentar e reduzir a escala na horizontal automaticamente. Uma política de escalabilidade com monitoramento do objetivo permite que

você selecione um resultado desejado e faça com que o grupo do Auto Scaling adicione e remova instâncias conforme necessário para atingir o resultado.

Para uma configuração de escalabilidade avançada, seu grupo do Auto Scaling pode ter mais de uma política de escalabilidade. Por exemplo, você pode definir uma ou mais políticas de escalabilidade de rastreamento de destino, uma ou mais políticas de escalabilidade em etapas, ou ambas. Isso fornece maior flexibilidade para abranger vários cenários.

Para ilustrar como várias políticas de escalabilidade dinâmica trabalham em conjunto, considere uma aplicação que use um grupo do Auto Scaling e uma fila do Amazon SQS para enviar solicitações a uma única instância do EC2. Para ajudar a garantir que a aplicação seja executada em níveis ideais, há duas políticas que controlam quando o grupo do Auto Scaling deve ter a escala aumentada na horizontal. Uma é uma política de rastreamento de destino que usa uma métrica personalizada para adicionar e remover capacidade com base no número de mensagens do SQS na fila. A outra é uma política de escalabilidade por etapas que usa a CloudWatch `CPUUtilization` métrica da Amazon para adicionar capacidade quando a instância excede 90% de utilização por um período de tempo especificado.

Quando há várias políticas em vigor ao mesmo tempo, há uma chance de que cada política instrua o grupo do Auto Scaling a ter a escala na horizontal aumentada (ou reduzida) ao mesmo tempo. Por exemplo, é possível que a `CPUUtilization` métrica aumente e ultrapasse o limite do CloudWatch alarme ao mesmo tempo em que a métrica personalizada do SQS aumente e ultrapasse o limite do alarme métrico personalizado.

Quando ocorrem essas situações, o Amazon EC2 Auto Scaling escolhe a política que fornece a maior capacidade para aumentar e para reduzir a escala na horizontal. Por exemplo, suponha que a política `CPUUtilization` execute uma instância, enquanto a política da fila do SQS executa duas instâncias. Se os critérios de aumento de escala na horizontal das duas políticas forem atendidos ao mesmo tempo, o Amazon EC2 Auto Scaling dará precedência à política da fila do SQS. Isso resulta na execução de duas instâncias no grupo do Auto Scaling.

A abordagem de dar precedência à política que fornece a maior capacidade se aplica mesmo quando as políticas usam critérios diferentes para aumentar. Por exemplo, se uma política terminar três instâncias, outra política diminuir o número de instâncias em 25%, e o grupo tiver oito instâncias no momento da redução da escala na horizontal, o Amazon EC2 Auto Scaling dará precedência à política que fornece o maior número de instâncias para o grupo. Isso faz com que o grupo do Auto Scaling termine duas instâncias (25% de 8 = 2). A intenção é evitar que o Amazon EC2 Auto Scaling remova instâncias demais.

No entanto, recomendamos cautela ao usar políticas de escalabilidade de rastreamento de destino com políticas de escalabilidade de etapas, pois conflitos entre essas políticas podem causar um comportamento indesejável. Por exemplo, se a política de escalabilidade de etapas iniciar uma atividade de redução antes que a política de rastreamento de destino esteja pronta para ser reduzida, a atividade de redução não será bloqueada. Após a conclusão da atividade de redução, a política de rastreamento de destino poderá instruir o grupo a expandir novamente.

Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling

Uma política de escalabilidade de rastreamento de destinos escala automaticamente a aplicação com base em um valor de métrica de destino. Ele se adapta automaticamente aos padrões de uso exclusivos de suas aplicações individuais. Isso permite que a aplicação mantenha uma performance ideal e alta utilização de suas instâncias do EC2 para obter eficiência de custos sem necessidade de intervenção manual.

Com o rastreamento de destinos, você seleciona uma métrica e um valor de destino para representar a utilização média ideal ou o nível de throughput para a aplicação. O Amazon EC2 Auto Scaling cria e CloudWatch gerencia os alarmes que invocam eventos de escalabilidade quando a métrica se desvia da meta. Isso é semelhante a como um termostato mantém a temperatura desejada.

Por exemplo, digamos que você tenha uma aplicação que seja executada em duas instâncias e queira que a utilização de CPU do grupo do Auto Scaling permaneça em cerca de 50% quando a carga na aplicação mudar. Isso fornece capacidade extra para lidar com picos de tráfego sem manter um número excessivo de recursos ociosos.

Você pode satisfazer essa necessidade criando uma política de escalabilidade com monitoramento de objetivo visando uma utilização média de 50% da CPU. Em seguida, o Application Auto Scaling aumentará a escala horizontalmente (aumento da capacidade) quando a CPU exceder 50% para lidar com o aumento de carga. Ele reduzirá a escala horizontalmente (diminuição da capacidade) quando a CPU estiver abaixo de 50% para otimizar os custos durante os períodos de baixa utilização.

Tópicos

- Várias políticas de escalabilidade de monitoramento de objetivo
- Escolher métricas
- Definir valor de objetivo
- Definir o tempo de aquecimento da instância

- Considerações
- Criar uma política de dimensionamento com monitoramento do objetivo
- Crie uma política de rastreamento de destino usando métricas de alta resolução para uma resposta mais rápida
- Criação de uma política de escalabilidade com monitoramento de destino com a matemática em métricas

Várias políticas de escalabilidade de monitoramento de objetivo

Para ajudar a otimizar o desempenho de escalonamento, você pode usar várias políticas de escalabilidade com monitoramento de objetivo juntas desde que cada uma delas use uma métrica diferente. Por exemplo, utilização e throughput podem promover influência cruzada. Sempre que uma dessas métricas muda, geralmente isso significa que outras métricas também serão afetadas. Portanto, o uso de várias métricas fornece informações adicionais sobre a carga em seu grupo do Auto Scaling e melhora a tomada de decisões ao determinar quanta capacidade adicionar ao seu grupo. Isso pode ajudar o Amazon EC2 Auto Scaling a tomar decisões mais informadas ao determinar a capacidade a ser adicionada ao seu grupo.

A intenção do Amazon EC2 Auto Scaling é sempre priorizar a disponibilidade. O grupo do Auto Scaling será aumentado a escala horizontalmente se qualquer uma das políticas de monitoramento do destino estiver pronta para ser aumentado a escala horizontalmente. Ele será reduzido a escala horizontalmente se todas as políticas de monitoramento do destino (com a parte de redução da escala horizontalmente ativada) estiverem prontas para serem reduzidas a escala horizontalmente.

Escolher métricas

É possível criar políticas de escalabilidade de rastreamento de destino com métricas predefinidas ou personalizadas. As métricas predefinidas facilitam o acesso às métricas mais usadas para escalabilidade. As métricas personalizadas permitem que você escale outras CloudWatch métricas disponíveis, incluindo métricas de alta resolução que são publicadas em intervalos mais precisos na ordem de alguns segundos. É possível publicar suas próprias métricas de alta resolução ou métricas publicadas por outros serviços da AWS .

Para obter mais informações sobre a criação de políticas de rastreamento de alvos usando métricas de alta resolução, consulte Crie uma política de rastreamento de destino usando métricas de alta resolução para uma resposta mais rápida.

O rastreamento de destino oferece suporte às métricas predefinidas a seguir:

- `ASGAverageCPUUtilization`: média de utilização da CPU do grupo do Auto Scaling.
- `ASGAverageNetworkIn`: número médio de bytes recebidos em todas as interfaces de rede pelo grupo do Auto Scaling.
- `ASGAverageNetworkOut`: número médio de bytes enviados em todas as interfaces de rede pelo grupo do Auto Scaling.
- `ALBRequestCountPerTarget`: número médio de solicitações do Application Load Balancer por destino para seu grupo do Auto Scaling.

> **Important**
>
> Outras informações valiosas sobre as métricas de utilização da CPU, E/S de rede e contagem de solicitações do Application Load Balancer por destino podem ser encontradas no tópico Listar as métricas CloudWatch disponíveis para suas instâncias no Guia do usuário do Amazon EC2 e CloudWatch as métricas do seu tópico Application Load Balancer no Guia do usuário para Application Load Balancers, respectivamente.

Você pode escolher outras CloudWatch métricas disponíveis ou suas próprias métricas CloudWatch especificando uma métrica personalizada. Para obter um exemplo que especifica uma especificação métrica personalizada para uma política de escalabilidade de rastreamento de metas usando o AWS CLI, consulte. Exemplo de políticas de escalabilidade para o AWS CLI

Lembre-se do seguinte ao escolher uma métrica:

- Recomendamos que você somente use as métricas que estejam disponíveis em intervalos de um minuto ou menos para ajudar a escalar mais rapidamente em resposta a alterações na utilização. Métricas publicadas em intervalos menores permitem que a política de rastreamento de metas detecte e responda mais rapidamente às mudanças na utilização do seu grupo do Auto Scaling.
- Se você escolher métricas predefinidas que sejam publicadas pelo Amazon EC2, como utilização de CPU, recomendamos que você habilite o monitoramento detalhado. Por padrão, todas as métricas do Amazon EC2 são publicadas em intervalos de cinco minutos, mas podem ser configuradas para um intervalo menor, de um minuto, ativando o monitoramento detalhado. Para obter informações sobre como habilitar o monitoramento detalhado, consulte Configurar monitoramento para instâncias do Auto Scaling.
- Nem todas as métricas personalizadas funcionam para rastreamento de destino. A métrica deve ser de utilização válida e descrever o quão ocupada uma instância está. O valor da métrica deve

aumentar e diminuir em proporção ao número das instâncias no grupo do Auto Scaling. Isso é para que os dados da métrica possam ser usados para expandir ou reduzir o número de instâncias. Por exemplo, a utilização da CPU de um grupo do Auto Scaling funcionará (ou seja, a métrica `CPUUtilization` do Amazon EC2 com a dimensão da métrica `AutoScalingGroupName`) se a carga no grupo do Auto Scaling for distribuída entre as instâncias.

- As métricas a seguir não funcionam para rastreamento de destino:
  - O número de solicitações recebidas pelo balanceador de carga voltadas para o grupo do Auto Scaling (ou seja, a métrica `RequestCount` do Elastic Load Balancing). O número de solicitações recebidas pelo balanceador de carga não é alterado com base na utilização do grupo do Auto Scaling.
  - A latência da solicitação do balanceador de carga (ou seja, a métrica `Latency` do Elastic Load Balancing). A latência da solicitação pode aumentar com base no aumento da utilização, mas não necessariamente muda de forma proporcional.
  - A CloudWatch métrica de fila do Amazon SQS. `ApproximateNumberOfMessagesVisible` O número de mensagens em uma fila pode não mudar proporcionalmente ao tamanho do grupo do Auto Scaling que processa mensagens da fila. Contudo, uma métrica personalizada que meça o número de mensagens na fila por instância do EC2 no grupo do Auto Scaling pode funcionar. Para obter mais informações, consulte Ajuste de escala baseado no Amazon SQS.
- Para usar a métrica `ALBRequestCountPerTarget`, é necessário especificar o parâmetro `ResourceLabel` a fim de identificar o grupo de destino do balanceador de carga que está associado à métrica. Para obter um exemplo que especifica o `ResourceLabel` parâmetro para uma política de escalabilidade de rastreamento de metas usando o AWS CLI, consulte. Exemplo de políticas de escalabilidade para o AWS CLI
- Quando uma métrica emite valores reais de 0 para CloudWatch (por exemplo,`ALBRequestCountPerTarget`), um grupo de Auto Scaling pode escalar até 0 quando não há tráfego para seu aplicativo por um período prolongado. A capacidade mínima do grupo deve estar definida como 0 para que seu grupo do Auto Scaling reduza a escala horizontalmente para 0 quando não houver solicitação roteada para ele.
- Em vez de publicar novas métricas para usar em sua política de escalabilidade, é possível usar a matemática métrica para combinar métricas existentes. Para obter mais informações, consulte Criação de uma política de escalabilidade com monitoramento de destino com a matemática em métricas.

Definir valor de objetivo

Ao criar uma política de escalabilidade com monitoramento de objetivo, é necessário especificar um valor para o objetivo. O valor-alvo representa o uso ou o throughput médio ideal para o grupo do Auto Scaling. Para usar os recursos de maneira econômica, defina o valor do objetivo com o número mais alto possível considerando um buffer razoável para aumentos inesperados de tráfego. Quando sua aplicação aumentar a escala horizontalmente para um fluxo de tráfego normal, o valor efetivo da métrica deve estar no valor desejado ou logo abaixo dele.

Quando uma política de dimensionamento é baseada no throughput, como o número de solicitações por destino para um Application Load Balancer, E/S de rede ou outras métricas de contagem, o valor-alvo representa o throughput médio ideal de uma única instância em um período de um minuto.

Definir o tempo de aquecimento da instância

Como opção, você pode especificar o número de segundos necessários para o aquecimento de uma instância recém-ativada. Até que o tempo de aquecimento especificado expire, uma instância não é contada para as métricas de instância do EC2 agregadas do grupo do Auto Scaling.

Enquanto as instâncias estiverem no período de aquecimento, suas políticas de escalabilidade somente aumentarão a escala na horizontal se o valor da métrica das instâncias que não estão se aquecendo for maior do que a utilização de destino da política.

Se o grupo voltar a aumentar a escala na horizontal, as instâncias que ainda estão se aquecendo serão contadas como parte da capacidade desejada para a próxima ação de aumento da escala na horizontal. A intenção é expandir de forma contínua (mas não excessivamente).

Enquanto a atividade de aumentar a escala na horizontal estiver em andamento, todas as atividades de reduzir a escala na horizontal iniciadas por políticas de escalabilidade serão bloqueadas até que as instâncias terminem de aquecer. Quando as instâncias terminarem de se aquecer, se ocorrer um evento de reduzir a escala horizontalmente, todas as instâncias atualmente em processo de encerramento serão contabilizadas na capacidade atual do grupo ao calcular a nova capacidade desejada. Portanto, não removemos mais instâncias do que o necessário do grupo do Auto Scaling.

Valor padrão

Se nenhum valor for definido, a política de escalabilidade usará o valor padrão, que é o valor para o aquecimento da instância padrão definido para o grupo. Se o aquecimento da instância padrão for nulo, ele retornará ao valor de desaquecimento padrão. Recomendamos usar o aquecimento padrão

da instância para facilitar a atualização de todas as políticas de escalabilidade quando o horário de aquecimento mudar.

Considerações

As considerações a seguir são aplicáveis ao trabalhar com políticas de escalabilidade com monitoramento de objetivo:

- Não crie, edite ou exclua os CloudWatch alarmes usados com uma política de escalabilidade de rastreamento de metas. O Amazon EC2 Auto Scaling cria e CloudWatch gerencia os alarmes associados às suas políticas de escalabilidade de rastreamento de destino e pode editá-los, substituí-los ou excluí-los quando necessário para personalizar a experiência de escalabilidade de seus aplicativos e seus padrões de utilização variáveis.
- Uma política de escalonamento com monitoramento de objetivo prioriza a disponibilidade durante períodos de níveis flutuantes de tráfego, reduzindo a escala na horizontal de maneira mais gradual quando o tráfego está diminuindo. Se você quiser maior controle, uma política de escalabilidade por etapas pode ser a melhor opção. É possível desabilitar temporariamente a parte de redução de uma política com rastreamento de destino. Isso ajuda a manter um número mínimo de instâncias para implantações com êxito.
- Se faltarem pontos de dados na métrica, isso fará com que o estado do CloudWatch alarme mude para`INSUFFICIENT_DATA`. Quando isso acontece, o Amazon EC2 Auto Scaling não poderá escalar seu grupo até que novos pontos de dados sejam encontrados.
- A matemática métrica pode ser útil se a métrica for intencionalmente relatada de maneira esparsa. Por exemplo, para usar os valores mais recentes, use a função `FILL(m1,REPEAT)`, na qual `m1` é a métrica.
- É possível ver lacunas entre o valor de destino e os pontos de dados de métrica reais. Isso ocorre porque agimos de maneira conservadora arredondando para cima ou para baixo, ao determinarmos quantas instâncias adicionar ou remover. Isso evita a adição de um número insuficiente de instâncias ou remova muitas instâncias. No entanto, para grupos do Auto Scaling menores, com um número menor de instâncias, a utilização do grupo pode parecer distante do valor do objetivo.

  Por exemplo, suponhamos que você defina um valor de objetivo de 50% para a utilização da CPU, e o seu grupo do Auto Scaling exceda o objetivo. Podemos determinar que a adição de 1,5 instância diminuirá a utilização da CPU em cerca de 50%. Como não é possível adicionar 1,5 instância, arredondamos para cima e adicionamos duas instâncias. Isso pode diminuir a utilização da CPU para um valor abaixo de 50%, mas garante que sua aplicação tenha recursos suficientes

para oferecer suporte a ele. Da mesma forma, se determinarmos que a remoção de 0,5 instâncias aumenta a utilização da CPU para mais de 50%, optaremos por não aumentar a escala até que a métrica diminua o suficiente para que pensemos que a escalabilidade não causará oscilação.

Para grupos do Auto Scaling maiores, com mais instâncias, a utilização é distribuída entre um maior número de instâncias, caso em que adicionar ou remover instâncias causa menos de uma lacuna entre o valor do objetivo e os pontos de dados de métrica reais.

- Uma política de escalabilidade com monitoramento do objetivo pressupõe que ela deve aumentar a escalabilidade de seu grupo do Auto Scaling quando a métrica especificada estiver acima do valor do objetivo. Você não pode usar uma política de escalabilidade com monitoramento do objetivo para aumentar horizontalmente a escala do seu grupo do Auto Scaling quando a métrica especificada estiver abaixo do valor do objetivo.

Criar uma política de dimensionamento com monitoramento do objetivo

Para criar uma política de escala de monitoramento do destino para o grupo do Auto Scaling, use um dos métodos a seguir.

Antes de começar, confirme se sua métrica preferida está disponível em intervalos de 1 minuto (em comparação com o intervalo padrão de 5 minutos das métricas do Amazon EC2).

Console

Para criar uma política de escalabilidade com monitoramento do objetivo para um grupo do Auto Scaling novo

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Selecione Criar grupo do Auto Scaling.
3. Nas etapas 1, 2 e 3, escolha as opções conforme desejado e prossiga para a Etapa 4: Configurar políticas de escalabilidade e tamanho do grupo.
4. Em Escalabilidade, especifique o intervalo no qual você deseja escalar atualizando a capacidade mínima desejada e a capacidade máxima desejada. Essas duas configurações permitem que seu grupo do Auto Scaling seja escalado dinamicamente. Para obter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.
5. Em Escalabilidade automática, escolha Política de escalabilidade com rastreamento do destino.

6. Para definir a política, faça o seguinte:

   a. Especifique um nome para a política.

   b. Escolha uma métrica para o Tipo de métrica.

      Se tiver escolhido Application Load Balancer request count per target (Contagem de solicitações do Application Load Balancer por destino), escolha um grupo de destino em Target group (Grupo de destino).

   c. Especifique um Target value (Valor de destino) para a métrica.

   d. (Opcional) Em Aquecimento das instâncias, atualize o valor de aquecimento de instância conforme necessário.

   e. (Opcional) Selecione Disable scale in to create only a scale-out policy (Desabilitar redução para criar somente uma política de expansão). Isso permite que você crie uma política de redução separada de um tipo diferente, se desejado.

7. Prossiga para criar o grupo do Auto Scaling. Sua política de escalabilidade será criada depois que o grupo do Auto Scaling for criado.

Para criar uma política de escalabilidade com monitoramento do objetivo para um grupo do Auto Scaling existente

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Verificar se os limites de escalabilidade estão definidos adequadamente. Por exemplo, se sua capacidade desejada já estiver no máximo, especifique um novo máximo para aumentar a escala horizontalmente. Para obter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.
4. Na guia Automatic scaling (Escalabilidade automática), em Dynamic scaling policies (Políticas dinâmicas de escalabilidade), selecione Create dynamic scaling policy (Criar política dinâmica de escalabilidade).
5. Para definir a política, faça o seguinte:

   a. Em Tipo de política, mantenha o padrão de Escalabilidade de rastreamento de destino.

   b. Especifique um nome para a política.

c. Escolha uma métrica para o Tipo de métrica. É possível escolher apenas um tipo de métrica. Para usar mais de uma métrica, crie várias políticas.

   Se você escolheu Application Load Balancer request count per target (Contagem de solicitações do balanceador de carga da aplicação por destino), escolha um grupo de destino em Target group (Grupo de destino).

d. Especifique um Target value (Valor de destino) para a métrica.

e. (Opcional) Em Aquecimento das instâncias, atualize o valor de aquecimento de instância conforme necessário.

f. (Opcional) Selecione Disable scale in to create only a scale-out policy (Desabilitar redução para criar somente uma política de expansão). Isso permite que você crie uma política de redução separada de um tipo diferente, se desejado.

6. Escolha Criar.

AWS CLI

Para criar uma política de escalabilidade de rastreamento de destino, é possível usar o exemplo a seguir para ajudar a começar. Substitua cada *`user input placeholder`* por suas próprias informações.

> **Note**
>
> Para obter mais exemplos, consulte Exemplo de políticas de escalabilidade para o AWS CLI.

Para criar uma política de escalabilidade com rastreamento do destino (AWS CLI)

1. Use o comando `cat` a seguir para especificar um valor de destino para sua política de escalabilidade e uma especificação de métrica personalizada em um arquivo JSON chamado `config.json` em seu diretório inicial. Veja a seguir um exemplo de configuração de rastreamento de destino que mantém a utilização média da CPU em 50%.

```
$ cat ~/config.json
{
  "TargetValue": 50.0,
  "PredefinedMetricSpecification":
    {
```

```
      "PredefinedMetricType": "ASGAverageCPUUtilization"
    }
}
```

Para obter mais informações, consulte a PredefinedMetricSpecificationReferência da API Amazon EC2 Auto Scaling.

2. Use o comando put-scaling-policy, juntamente com o arquivo `config.json` criado na etapa anterior, para criar sua política de escalabilidade.

```
aws autoscaling put-scaling-policy --policy-name cpu50-target-tracking-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
```

Se for bem-sucedido, esse comando retornará os nomes ARNs e os dois CloudWatch alarmes criados em seu nome.

```
{
    "PolicyARN": "arn:aws:autoscaling:us-
west-2:123456789012:scalingPolicy:228f02c2-c665-4bfd-
aaac-8b04080bea3c:autoScalingGroupName/my-asg:policyName/cpu50-target-tracking-
scaling-policy",
    "Alarms": [
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e",
            "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
        },
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
            "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
        }
    ]
}
```

Crie uma política de rastreamento de destino usando métricas de alta resolução para uma resposta mais rápida

O rastreamento de metas oferece suporte a CloudWatch métricas de alta resolução com pontos de dados de segundo nível que são publicados em intervalos menores que um minuto. Configure políticas de rastreamento de metas para monitorar a utilização por meio de CloudWatch métricas de alta resolução para aplicativos que têm padrões de demanda voláteis, como APIs de atendimento ao cliente, serviços de streaming ao vivo, sites de comércio eletrônico e processamento de dados sob demanda. Para obter maior precisão na correspondência entre capacidade e demanda, o rastreamento de destino usa esse monitoramento refinado para detectar e responder às mudanças na demanda e na utilização de suas instâncias do EC2 com mais rapidez.

Para obter mais informações sobre como publicar suas métricas em alta resolução, consulte Publicar métricas personalizadas no Guia do CloudWatch usuário da Amazon. Para acessar e publicar métricas do EC2, como a utilização da CPU em alta resolução, talvez você queira usar CloudWatch o agente.

Regiões da AWS

O rastreamento de alvos usando métricas de alta resolução está disponível em todos Regiões da AWS , exceto no AWS GovCloud (US) Regions.

Como funciona a política de rastreamento de destino com métricas de alta resolução

Você cria políticas de rastreamento de destino definindo a métrica que deseja rastrear e o valor de destino que deseja manter para a métrica. Para escalar uma métrica de alta resolução, especifique o nome da métrica e defina o período métrico no qual o rastreamento de destino observa essa métrica a um valor inferior a 60 segundos. Atualmente, o intervalo mais baixo com suporte é de 10 segundos. É possível publicar sua métrica em intervalos menores do que isso.

> **Note**
>
> Não há suporte a um período métrico maior que 60.

Você pode configurar o rastreamento de metas em uma única CloudWatch métrica ou consultar várias CloudWatch métricas e usar expressões matemáticas para criar novas séries temporais únicas com base nessas métricas. Ambas as opções permitem que você defina o período métrico.

Exemplos

Exemplo 1

O exemplo a seguir cria uma política de rastreamento de metas com base em uma CloudWatch métrica de alta resolução. A métrica é publicada com resolução de 10 segundos. Ao definir o período, é possível habilitar o rastreamento de metas para monitorar essa métrica com uma granularidade de 10 segundos. Substitua cada *user input placeholder* por suas próprias informações.

```
$ cat ~/config.json
{
  "TargetValue": 100.0,
  "CustomizedMetricSpecification": {
      "MetricName": "MyHighResolutionMetric",
      "Namespace": "MyNamespace",
      "Dimensions": [
        {
          "Name": "MyOptionalDimensionName",
          "Value": "MyOptionalMetricDimensionValue"
        }
      ],
      "Statistic": "Average",
      "Unit": "None"
      "Period": "10
  }
}
```

Exemplo 2

É possível usar expressões matemáticas métricas para combinar várias métricas em uma única série temporal para escalabilidade. A matemática métrica é particularmente útil para converter métricas existentes em média por instância. A conversão de métricas é essencial porque o rastreamento de destino pressupõe que a métrica seja inversamente proporcional à capacidade do grupo do Auto Scaling. Portanto, quando a capacidade aumenta, a métrica deve diminuir quase na mesma proporção.

Por exemplo, suponha que você tenha uma métrica que represente os trabalhos pendentes a serem processados pela aplicação. É possível usar matemática métrica para dividir os trabalhos pendentes pela capacidade de execução do grupo do Auto Scaling. O Auto Scaling publica a métrica de capacidade com granularidade de 1 minuto. Portanto, não haverá nenhum valor para essa métrica

em intervalos de menos de um minuto. Se você quiser usar uma resolução mais alta para escalar, isso pode levar a uma incompatibilidade de período entre a capacidade e a métrica de trabalho pendente. Para evitar essa incompatibilidade, recomendamos que você use a expressão FILL para preencher os valores ausentes com o número da capacidade registrado no carimbo de data/hora do minuto anterior.

O exemplo a seguir usa matemática métrica para dividir a métrica de trabalhos pendentes pela capacidade. Por um período, estamos definindo as duas métricas em 10 segundos. Como a métrica é publicada em intervalos de 1 minuto, estamos usando a operação FILL na métrica de capacidade.

Para usar matemática métrica para modificar várias métricas

```
{
    "CustomizedMetricSpecification": {
        "Metrics": [
            {
                "Label": "Pending jobs to be processed",
                "Id": "m1",
                "MetricStat": {
                    "Metric": {
                        "MetricName": "MyPendingJobsMetric",
                        "Namespace": "Custom",
                    },
                    "Stat": "Sum"
                    "Period": 10
                },
                "ReturnData": false
            },
            {
                "Label": "Get the running instance capacity (matching the period to
 that of the m1)",
                "Id": "m2",
                "MetricStat": {
                    "Metric": {
                        "MetricName": "GroupInServiceInstances",
                        "Namespace": "AWS/AutoScaling",
                        "Dimensions": [
                            {
                                "Name": "AutoScalingGroupName",
                                "Value": "my-asg"
                            }
                        ]
                    },
```

```
                    "Stat": "Average"
                    "Period": 10
                },
                "ReturnData": false
            },
            {
                "Label": "Calculate the pending job per capacity (note the use of the
 FILL expression)",
                "Id": "e1",
                "Expression": "m1 / FILL(m2,REPEAT)",
                "ReturnData": true
            }
        ]
    },
    "TargetValue": 100
}
```

Considerações

Considere os fatores a seguir ao usar rastreamento de destino e métricas de alta resolução.

- Para garantir que você não tenha pontos de dados ausentes que possam levar a resultados de escalonamento automático indesejados, sua CloudWatch métrica deve ser publicada com a mesma resolução ou maior do que o período especificado.
- Defina o valor alvo como o valor per-instance-per-minute métrico que você deseja manter para seu grupo de Auto Scaling. Definir um valor de destino apropriado é crucial se você usar uma métrica cujo valor pode se multiplicar com base no período da métrica. Por exemplo, qualquer métrica baseada em contagem, como contagens de solicitações ou trabalhos pendentes que usem a estatística SUM, terá um valor métrico diferente dependendo do período escolhido. Você ainda deve presumir que está estabelecendo uma meta em relação à média por minuto.
- Embora não haja taxas adicionais para usar o Amazon EC2 Auto Scaling, você deve pagar pelos recursos, como CloudWatch instâncias, métricas e alarmes do Amazon EC2. CloudWatch Os alarmes de alta resolução criados no exemplo anterior têm preços diferentes dos alarmes padrão. CloudWatch Para obter mais informações sobre CloudWatch preços, consulte Amazon CloudWatch Pricing.
- O rastreamento de destino exige que as métricas representem a utilização média por instância de suas instâncias do EC2. Para conseguir isso, é possível usar operações matemáticas métricas como parte da configuração da política de rastreamento de destino. Divida sua métrica pela capacidade de execução do seu grupo do Auto Scaling. Certifique-se de que o mesmo período

métrico seja definido para cada uma das métricas que você usar para criar uma única série temporal. Se essas métricas forem publicadas em intervalos diferentes, use a operação FILL na métrica com o intervalo mais alto para preencher os pontos de dados ausentes.

Criação de uma política de escalabilidade com monitoramento de destino com a matemática em métricas

Usando a matemática métrica, você pode consultar várias CloudWatch métricas e usar expressões matemáticas para criar novas séries temporais com base nessas métricas. Você pode visualizar as séries temporais resultantes no CloudWatch console e adicioná-las aos painéis. Para obter mais informações sobre matemática métrica, consulte Usando matemática métrica no Guia CloudWatch do usuário da Amazon.

As considerações a seguir se aplicam a expressões matemática em métricas:

- Você pode consultar qualquer CloudWatch métrica disponível. Cada métrica corresponde a uma combinação exclusiva de nome de métrica, espaço nominal e zero ou mais dimensões.
- Você pode usar qualquer operador aritmético (+ - */^), função estatística (como AVG ou SUM) ou outra função compatível. CloudWatch
- Você pode usar as métricas e os resultados de outras expressões matemáticas nas fórmulas da expressão matemática.
- Qualquer expressão usada em uma especificação de métrica deve eventualmente retornar uma única série temporal.
- Você pode verificar se uma expressão matemática métrica é válida usando o CloudWatch console ou a CloudWatch GetMetricDataAPI.

Exemplo: lista de pendências da fila do Amazon SQS por instância

Para calcular a lista de pendências da fila do Amazon SQS por instância, use o número aproximado de mensagens disponíveis para recuperação da fila e divida esse número pela capacidade de execução do grupo do, que corresponde ao número de instâncias no estado`InService`. Para obter mais informações, consulte Ajuste de escala baseado no Amazon SQS.

A lógica da expressão é a seguinte:

```
sum of (number of messages in the queue)/(number of InService instances)
```

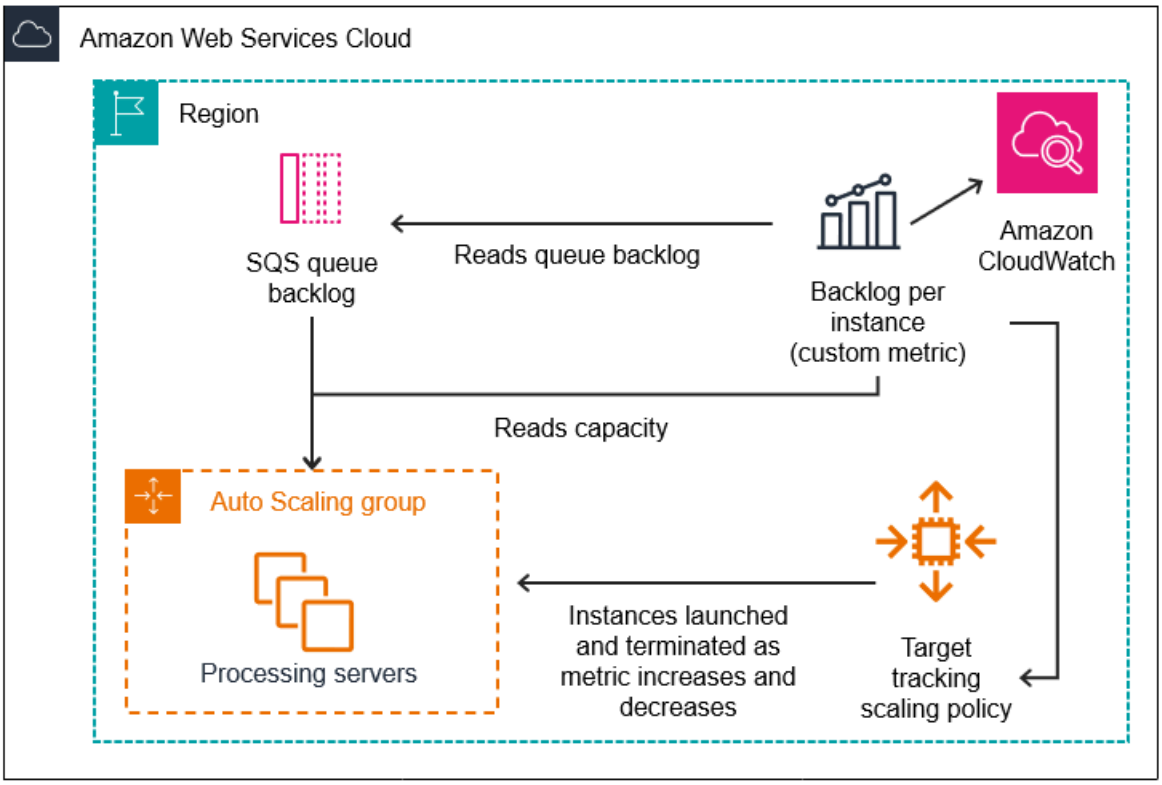
Então, suas informações CloudWatch métricas são as seguintes.

ID	CloudWatch métrica	Estatística	Período
m1	ApproximateNumberOfMessagesVisible	Soma	1 minuto
m2	GroupInServiceInstances	Média	1 minuto

O ID e a expressão matemáticos da métrica são os seguintes:

ID	Expressão
e1	`(m1)/(m2)`

O diagrama a seguir ilustra a arquitetura desse padrão.



Para usar essa matemática em métricas na criação de uma política de escalabilidade com monitoramento de destino (AWS CLI)

1. Armazene a expressão matemática em métricas como parte de uma especificação de métrica personalizada em um arquivo JSON denominado `config.json`.

   Use o exemplo a seguir como auxílio para começar. Substitua cada *`user input placeholder`* por suas próprias informações.

```
{
    "CustomizedMetricSpecification": {
        "Metrics": [
            {
                "Label": "Get the queue size (the number of messages waiting to be
 processed)",
                "Id": "m1",
                "MetricStat": {
                    "Metric": {
                        "MetricName": "ApproximateNumberOfMessagesVisible",
                        "Namespace": "AWS/SQS",
                        "Dimensions": [
                            {
                                "Name": "QueueName",
                                "Value": "my-queue"
                            }
                        ]
                    },
                    "Stat": "Sum"
                },
                "ReturnData": false
            },
            {
                "Label": "Get the group size (the number of InService instances)",
                "Id": "m2",
                "MetricStat": {
                    "Metric": {
                        "MetricName": "GroupInServiceInstances",
                        "Namespace": "AWS/AutoScaling",
                        "Dimensions": [
                            {
                                "Name": "AutoScalingGroupName",
                                "Value": "my-asg"
                            }
```

```
                        ]
                    },
                    "Stat": "Average"
                },
                "ReturnData": false
            },
            {
                "Label": "Calculate the backlog per instance",
                "Id": "e1",
                "Expression": "m1 / m2",
                "ReturnData": true
            }
        ]
    },
    "TargetValue": 100
}
```

Para obter mais informações, consulte a TargetTrackingConfigurationReferência da API Amazon EC2 Auto Scaling.

> **Note**
>
> Veja a seguir alguns recursos adicionais que podem ajudá-lo a encontrar nomes de métricas, namespaces, dimensões e estatísticas para CloudWatch métricas:
>
> - Para obter informações sobre as métricas disponíveis para AWS serviços, consulte AWS serviços que publicam CloudWatch métricas no Guia CloudWatch do usuário da Amazon.
> - Para obter o nome exato da métrica, o namespace e as dimensões (se aplicável) de uma CloudWatch métrica com o AWS CLI, consulte list-metrics.

2. Para criar essa política, execute o put-scaling-policycomando usando o arquivo JSON como entrada, conforme demonstrado no exemplo a seguir.

```
aws autoscaling put-scaling-policy --policy-name sqs-backlog-target-tracking-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
```

Se for bem-sucedido, esse comando retornará o Amazon Resource Name (ARN) da política e o ARNs dos dois CloudWatch alarmes criados em seu nome.

```
{
    "PolicyARN": "arn:aws:autoscaling:us-
west-2:123456789012:scalingPolicy:228f02c2-c665-4bfd-
aaac-8b04080bea3c:autoScalingGroupName/my-asg:policyName/sqs-backlog-target-
tracking-scaling-policy",
    "Alarms": [
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e",
            "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
        },
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
            "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
        }
    ]
}
```

> **Note**
>
> Se esse comando gerar um erro, verifique se você atualizou o AWS CLI localmente para a versão mais recente.

Políticas de escalabilidade simples e em etapas do Amazon EC2 Auto Scaling

O escalonamento por etapas e as políticas de escalabilidade simples escalam a capacidade do seu grupo de Auto Scaling em incrementos predefinidos com base em alarmes. CloudWatch É possível definir políticas de escalabilidade separadas para lidar com o aumento horizontal da escala (aumento

da capacidade) e com a redução horizontal da escala (diminuição da capacidade) quando um limite de alarme é violado.

A capacidade do grupo do Auto Scaling é medida em termos de instâncias ou unidades de capacidade se você estiver usando pesos de instância. Além disso, há uma diferença entre a capacidade desejada e a capacidade atual.

- Capacidade desejada: o número de instâncias (ou unidades de capacidade) que você deseja ter em seu grupo. A capacidade desejada pode ser ajustada manual ou automaticamente usando políticas de escalabilidade.
- Capacidade atual: o número de instâncias (ou unidades de capacidade) em seu grupo que passaram pelos períodos de aquecimento e desaquecimento, estão em execução e prontas para serem usadas.

Com o escalonamento por etapas e o escalonamento simples, você cria e gerencia os CloudWatch alarmes que invocam o processo de escalabilidade. Quando um alarme é violado, o Amazon EC2 Auto Scaling inicia a política de escalabilidade associada a esse alarme.

É altamente recomendável usar uma política de escalabilidade de rastreamento de destino para escalar métricas, como utilização média de CPU ou contagem média de solicitações por destino. Métricas que diminuem quando a capacidade aumenta e aumentam quando a capacidade diminui podem ser usadas para expandir ou reduzir proporcionalmente o número de instâncias usando o rastreamento de destino. Isso ajuda a garantir que o Amazon EC2 Auto Scaling siga estritamente a curva de demanda para suas aplicações. Para obter mais informações, consulte Políticas de escalabilidade de rastreamento de destino.

Conteúdo

- Funcionamento das políticas de escalabilidade em etapas
- Ajustes em etapas para escalabilidade em etapas
- Tipos de ajuste da escalabilidade
- Aquecimento da instância
- Considerações
- Como criar uma política de escalabilidade em etapas para expansão
- Como criar uma política de escalabilidade em etapas para redução
- Políticas de escalabilidade simples

Funcionamento das políticas de escalabilidade em etapas

Para usar o escalonamento por etapas, primeiro você cria um CloudWatch alarme que monitora uma métrica para seu grupo de Auto Scaling. Defina a métrica, o valor limite e o número de períodos de avaliação que determinam uma violação de alarme. Em seguida, crie uma política de escala por etapas que defina como escalar seu grupo quando o limite de alarme for ultrapassado. É possível usar um percentual da capacidade atual do seu grupo do Auto Scaling ou unidades de capacidade para o tipo de ajuste de escalabilidade. Para obter mais informações, consulte Tipos de ajuste da escalabilidade.

Adicione os ajustes de etapas na política. É possível definir diferentes ajustes de etapas com base na dimensão da violação do alarme. Por exemplo:

- Aumentar a escala horizontalmente em 10 unidades de capacidade, se a métrica de alarme atingir 60%.
- Aumentar a escala horizontalmente em 30 unidades de capacidade, se a métrica de alarme atingir 75%.
- Aumentar a escala horizontalmente em 40 unidades de capacidade, se a métrica de alarme atingir 85%.

Quando o limite de alarme for violado durante o número especificado de períodos de avaliação, o Amazon EC2 Auto Scaling aplicará os ajustes de etapas definidos na política. Os ajustes podem continuar para violações de alarmes adicionais até que o estado do alarme retorne a OK.

Cada instância tem um período de aquecimento para evitar que as atividades de escala sejam muito reativas às mudanças que ocorrem em curtos períodos de tempo. Opcionalmente, é possível configurar os períodos de esfriamento para a política de escalabilidade. Recomendamos usar o aquecimento padrão da instância para facilitar a atualização de todas as políticas de escalabilidade quando o horário de aquecimento mudar. Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

As políticas de escala simples são semelhantes às políticas de escala por etapas, exceto pelo fato de que se baseiam em um único ajuste de escala, com um período de espera entre cada atividade de escala. Para obter mais informações, consulte Políticas de escalabilidade simples.

Ajustes em etapas para escalabilidade em etapas

Ao criar uma política de escalabilidade em etapas, especifique um ou mais ajustes em etapas que dimensionem automaticamente o número de instâncias de forma dinâmica com base no tamanho da violação do alarme. Cada ajuste em etapas especifica o seguinte:

- Um limite inferior para o valor da métrica
- Um limite superior para o valor da métrica
- O valor de acordo com o qual dimensionar com base no tipo de ajuste de dimensionamento

CloudWatch agrega pontos de dados métricos com base na estatística da métrica associada ao seu CloudWatch alarme. Quando o alarme é violado, a política de dimensionamento apropriada é invocada. O Amazon EC2 Auto Scaling aplica o tipo de agregação aos CloudWatch pontos de dados métricos mais recentes (em oposição aos dados métricos brutos). Ele compara esse valor de métrica agregada com os limites superior e inferior definidos pelo ajustes em etapa para determinar qual deles deve ser executado.

Você especifica os limites superior e inferior em relação ao limite de ruptura. Por exemplo, digamos que você tenha criado um CloudWatch alarme e uma política de expansão para quando a métrica estiver acima de 50%. Em seguida, você criou um segundo alarme e uma política para reduzir a escala horizontalmente em momentos em que a métrica está abaixo de 50%. Você tem um conjunto de ajustes em etapas com um tipo de ajuste `PercentChangeInCapacity` (ou Percent of group (Percentual de grupo) no console) para cada política:

Exemplo: ajustes em etapas para política de expansão

Limite inferior	Limite superior	Ajuste
0	10	0
10	20	10
20	nulo	30

Exemplo: ajustes em etapas para política de redução

Limite inferior	Limite superior	Ajuste
-10	0	0
-20	-10	-10
nulo	-20	-30

Isso cria a seguinte configuração de escalabilidade.

```
Metric value

-infinity          30%    40%          60%     70%             infinity
-------------------------------------------------------------------------
          -30%      | -10% | Unchanged  | +10%  |       +30%
-------------------------------------------------------------------------
```

Por exemplo, digamos que você tenha um grupo do Auto Scaling que tenha uma capacidade atual e uma capacidade desejada de 10. Os pontos a seguir resumem o comportamento da configuração de escalabilidade em relação às capacidades desejada e atual do grupo:

- A capacidade atual e desejada será mantida enquanto o valor agregado da métrica for maior que 40 e menor que 60.
- Se o valor da métrica chegar a 60, a capacidade desejada do grupo aumenta em 1 instância, para 11 instâncias, com base no segundo ajuste em etapas da política de expansão (adicionar 10% de 10 instâncias). Depois que a nova instância estiver em execução e seu tempo de aquecimento especificado expirar, a capacidade atual do grupo aumenta para 11 instâncias. Se o valor da métrica subir para 70 mesmo após esse aumento na capacidade, a capacidade desejada do grupo aumentará em outras 3 instâncias, para 14 instâncias. Isso se baseia no ajuste da terceira etapa da política de expansão (adicione 30% de 11 instâncias, 3,3 instâncias, arredondadas para 3 instâncias).
- Se o valor da métrica chegar a 40, a capacidade desejada do grupo será reduzida em 1 instância, para 13 instâncias, com base no segundo ajuste em etapas da política de redução (removerá 10% das 14 instâncias, 1,4 instâncias, arredondadas para 1 instância). Se o valor da métrica cair para 30 mesmo após essa diminuição na capacidade, a capacidade desejada do grupo diminuirá em

outras 3 instâncias, para 10 instâncias. Isso se baseia no ajuste da terceira etapa da política de expansão (remova 30% de 13 instâncias, 3,9 instâncias, arredondadas para 3 instâncias).

Ao especificar os ajustes em etapas para sua política de escalabilidade, observe o seguinte:

- Se você usar o Console de gerenciamento da AWS, você especifica os limites superior e inferior como valores absolutos. Se você usa o AWS CLI ou um SDK, especifica os limites superior e inferior em relação ao limite de violação.
- Os intervalos de seus ajustes em etapas não podem se sobrepor ou ter uma lacuna.
- Somente um ajuste em etapas pode ter um limite inferior nulo (infinito negativo). Se um ajuste em etapas tiver um limite inferior negativo, não deverá haver um ajuste em etapas com um limite inferior nulo.
- Somente um ajuste em etapas pode ter um limite superior nulo (infinito positivo). Se um ajuste em etapas tiver um limite superior positivo, deverá haver um ajuste em etapas com um limite superior nulo.
- Os limites inferior e superior não podem ser nulos no mesmo ajuste em etapas.
- Se o valor da métrica estiver acima do limite de violação, o limite inferior será inclusivo e o limite superior será exclusivo. Se o valor da métrica estiver abaixo do limite de violação, o limite inferior será exclusivo e o limite superior será inclusivo.

Tipos de ajuste da escalabilidade

É possível definir uma política de escalabilidade que execute a ação de escalabilidade ideal, com base no tipo de ajuste de escalabilidade escolhido. É possível especificar o tipo de ajuste como um percentual da capacidade atual do seu grupo do Auto Scaling ou em unidades de capacidade. Normalmente, uma unidade de capacidade significa uma instância, a menos que você esteja usando o recurso de peso da instância.

O Amazon EC2 Auto Scaling oferece suporte aos seguintes tipos de ajuste de escalabilidade simples e em etapa:

- `ChangeInCapacity`: aumentar ou diminuir a capacidade atual do grupo no valor especificado. Um valor de ajuste positivo aumenta a capacidade e um valor negativo diminui a capacidade. Por exemplo: se a capacidade atual do grupo for 3 e o ajuste for 5, quando essa política for executada, adicionaremos 5 unidades de capacidade à capacidade, para um total de 8 unidades de capacidade.

- `ExactCapacity`: alterar a capacidade atual do grupo para o valor especificado. Especifique um valor não negativo com esse tipo de ajuste. Exemplo: se a capacidade atual do grupo for 3 instâncias e o ajuste for 5, quando essa política for executada, alteraremos a capacidade para 5 unidades de capacidade.
- `PercentChangeInCapacity`: aumentar ou diminuir a capacidade atual do grupo no percentual especificado. Um valor positivo aumenta a capacidade e um valor negativo diminui a capacidade. Por exemplo: se a capacidade atual for 10 e o ajuste for 10%, quando essa política for executada, adicionaremos 1 unidade de capacidade à capacidade, para um total de 11 unidades de capacidade.

> **Note**
>
> Se o valor resultante não for um inteiro, o arredondamento é feito da seguinte forma:
>
> - Valores maiores que 1 serão arredondados para baixo. Por exemplo, `12.7` será arredondado para `12`.
> - Os valores entre 0 e 1 serão arredondados para 1. Por exemplo, `.67` será arredondado para `1`.
> - Os valores entre 0 e -1 serão arredondados para -1. Por exemplo, `-.58` será arredondado para `-1`.
> - Os valores menores que -1 serão arredondado para cima. Por exemplo, `-6.67` será arredondado para `-6`.

Com `PercentChangeInCapacity`, também é possível especificar o número mínimo de instâncias a serem dimensionadas usando o parâmetro `MinAdjustmentMagnitude`. Por exemplo, vamos supor que você crie uma política que adiciona 25% e especifique um incremento mínimo de 2 instâncias. Se você tiver um grupo do Auto Scaling com 4 instâncias e a política de escalabilidade for executada, 25% de 4 será 1 instância. No entanto, como você especificou um incremento mínimo de 2, serão adicionadas 2 instâncias.

Quando o peso de instâncias é usado, o efeito de definir o parâmetro `MinAdjustmentMagnitude` para um valor diferente de zero é alterado. O valor é em unidades de capacidade. Para definir o número mínimo de instâncias a serem escaladas, defina esse parâmetro para um valor que seja, pelo menos, tão grande quanto o maior peso da instância.

Se você estiver usando ponderação de instâncias, lembre-se de que a capacidade atual do seu grupo do Auto Scaling poderá exceder a capacidade desejada, conforme necessário. Se o seu

número absoluto para redução, ou o valor que a porcentagem informar para redução, for menor que a diferença entre a capacidade atual e a desejada, nenhuma ação de escalabilidade será executada. É necessário levar em conta esses comportamentos ao analisar o resultado de uma política de dimensionamento quando um limite de alarme é violado. Por exemplo, vamos supor que a capacidade desejada seja 30 e a capacidade atual seja 32. Quando o alarme é violado, se a política de dimensionamento diminuir a capacidade desejada em um, nenhuma ação de dimensionamento será realizada.

Aquecimento da instância

Para dimensionamento em etapas, você pode especificar o número de segundos necessários para o aquecimento de uma instância recém-iniciada. Até que o tempo de aquecimento especificado expire, uma instância não é contada para as métricas de instância do EC2 agregadas do grupo do Auto Scaling.

Enquanto as instâncias estiverem no período de aquecimento, suas políticas de escalabilidade somente aumentarão a escala na horizontal se o valor da métrica de instâncias que não estão aquecendo for maior do que o limite alto do alarme da política.

Se o grupo voltar a aumentar a escala na horizontal, as instâncias que ainda estão se aquecendo serão contadas como parte da capacidade desejada para a próxima ação de aumento da escala na horizontal. Portanto, várias violações de alarme que caem no intervalo do mesmo ajuste em etapas resultam em uma única ação de escalabilidade. A intenção é expandir de forma contínua (mas não excessivamente).

Por exemplo, vamos supor que você cria uma política com duas etapas. A primeira etapa adiciona 10 por cento quando a métrica chega a 60, e a segunda etapa adiciona 30 por cento quando a métrica chega a 70 por cento. Seu grupo do Auto Scaling tem uma capacidade desejada e atual de 10. A capacidade atual e desejada não altera enquanto o valor agregado da métrica for menor que 60. Suponha que a métrica chegue a 60, então 1 instância é adicionada (10 por cento de 10 instâncias). Em seguida, a métrica chega a 62 enquanto a nova instância ainda está se aquecendo. A política de escalabilidade calcula a nova capacidade desejada com base na capacidade atual, que ainda é 10. No entanto, a capacidade desejada do grupo já aumentou para 11 instâncias, portanto, a política de escalabilidade não aumenta mais a capacidade desejada. Se a métrica chegar a 70 enquanto a nova instância ainda está em processo de aquecimento, deveremos adicionar 3 instâncias (30% de 10 instâncias). No entanto, como a capacidade desejada do grupo já é 11, adicionaremos apenas 2 instâncias, para uma nova capacidade desejada de 13 instâncias.

Enquanto a atividade de aumentar a escala na horizontal estiver em andamento, todas as atividades de reduzir a escala na horizontal iniciadas por políticas de escalabilidade serão bloqueadas até que as instâncias terminem de aquecer. Quando as instâncias terminarem de se aquecer, se ocorrer um evento de reduzir a escala horizontalmente, todas as instâncias atualmente em processo de encerramento serão contabilizadas na capacidade atual do grupo ao calcular a nova capacidade desejada. Portanto, não removemos mais instâncias do que o necessário do grupo do Auto Scaling. Por exemplo, enquanto uma instância já estiver sendo encerrada, se um alarme estiver em violação no intervalo do mesmo ajuste de etapa que diminuiu a capacidade desejada em 1, nenhuma ação de escalabilidade será realizada.

Valor padrão

Se nenhum valor for definido, a política de escalabilidade usará o valor padrão, que é o valor para o aquecimento da instância padrão definido para o grupo. Se o aquecimento da instância padrão for nulo, ele retornará ao valor de desaquecimento padrão.

Considerações

As considerações a seguir são aplicáveis ao trabalhar com políticas de escalabilidade simples e em etapas:

- Avalie se é possível prever os ajustes em etapas na aplicação com precisão suficiente para usar a escalabilidade em etapas. Se a métrica de escalabilidade aumentar ou diminuir proporcionalmente à capacidade do destino dimensionável, recomendamos que você use uma política de escalabilidade de rastreamento do objetivo. Você ainda tem a opção de usar a escalabilidade em etapas como política adicional para uma configuração mais avançada. Por exemplo, é possível configurar uma resposta mais agressiva quando a utilização atinge determinado nível.
- Para evitar oscilações, certifique-se de escolher uma margem adequada entre os limites de redução e aumento da escala. Oscilação é um ciclo infinito de aumento e redução de escala horizontal. Ou seja, se o sistema adotar alguma ação de escalabilidade, o valor da métrica mudaria e iniciaria outra ação de escalabilidade na direção inversa.

Como criar uma política de escalabilidade em etapas para expansão

Para criar uma política de escala por etapas para o aumentar a escala horizontalmente do seu grupo do Auto Scaling, use um dos métodos a seguir:

Console

Etapa 1: criar um CloudWatch alarme para o limite máximo métrico

1. Abra o CloudWatch console em https://console.aws.amazon.com/cloudwatch/.
2. Se necessário, altere a região da . Na barra de navegação, selecione a região na qual o grupo do Auto Scaling reside.
3. No painel de navegação, escolha Alarms, All alarms (Alarmes, Todos os alarmes) e Create alarm (Criar alarme).
4. Escolha Selecionar métrica.
5. Na guia All metrics (Todas as métricas), escolha EC2, By Auto Scaling Group (Por grupo do Auto Scaling) e insira o nome do grupo do Auto Scaling no campo de pesquisa. Depois, selecione `CPUUtilization` e escolha Selecionar métrica. A página Especificar métrica e condições será exibida, mostrando um gráfico e outras informações sobre a métrica.
6. Em Period (Período), escolha o período de avaliação para o alarme, por exemplo, 1 minuto. Ao avaliar o alarme, todos os períodos são agregados em um único ponto de dados.

   > **Note**
   >
   > Um período mais curto cria um alarme mais sensível.

7. Em Condições, faça o seguinte:

   - Em Tipo de limite, escolha Estático.
   - Em Whenever **`CPUUtilization`** is (Sempre que for), especifique se você deseja que o valor da métrica seja maior que, maior que ou igual a, menor que, ou menor que ou igual ao limite de violação do alarme. Em than (que), insira o valor do limite desejado de violação de alarme.

8. Em Configuração adicional, faça o seguinte:

   - Em Datapoints to alarm (Pontos de dados para alarme), insira o número de pontos de dados (períodos de avaliação) durante os quais o valor da métrica deverá atender às condições de limite para o alarme. Por exemplo, com dois períodos consecutivos de 5 minutos, o estado de alarme levaria 10 minutos para ser invocado.
   - Em Tratamento de dados ausentes, escolha Tratar dados ausentes como inválidos (limite de violação). Para obter mais informações, consulte Configurando como CloudWatch os alarmes tratam dados perdidos no Guia CloudWatch do usuário da Amazon.

9. Escolha Próximo.

   A página Configure actions (Configurar ações) é exibida.

10. Em Notification (Notificação), selecione um tópico do Amazon SNS para notificar quando o alarme estiver no estado `ALARM`, `OK` ou `INSUFFICIENT_DATA`.

    Para que o alarme envie várias notificações para o mesmo estado de alarme ou para diferentes estados de alarme, escolha Add notification (Adicionar notificação).

    Para que o alarme não envie notificações, escolha Remove (Remover).

11. Você pode deixar vazias as outras seções da página Configure actions (Configurar ações). Deixar as outras seções vazias cria um alarme sem associá-lo a uma política de escalabilidade. Em seguida, você pode associar o alarme a uma política de escalabilidade do console do Amazon EC2 Auto Scaling.

12. Escolha Próximo.

13. Insira um nome (por exemplo, `Step-Scaling-AlarmHigh-AddCapacity`) e, opcionalmente, uma descrição para o alarme e escolha Próximo.

14. Selecione Criar alarme.

Use o procedimento a seguir para continuar de onde parou depois de criar o CloudWatch alarme.

Etapa 2: Criar uma política de escalabilidade para aumentar a escala horizontalmente

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.

2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.

3. Verificar se os limites de escalabilidade estão definidos adequadamente. Por exemplo, se sua capacidade desejada já estiver no máximo, especifique um novo máximo para aumentar a escala horizontalmente. Para obter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.

4. Na guia Automatic scaling (Escalabilidade automática), em Dynamic scaling policies (Políticas dinâmicas de escalabilidade), selecione Create dynamic scaling policy (Criar política dinâmica de escalabilidade).

5. Para Tipo de política, escolha Escala por etapas e especifique um nome para a política.

6. Para CloudWatch alarme, escolha seu alarme. Se você ainda não criou um alarme, escolha Criar um CloudWatch alarme e conclua as etapas 4 a 14 no procedimento anterior para criar um alarme.
7. Especifique a alteração no tamanho do grupo atual que essa política fará quando executada usando Take the action (Executar a ação). É possível adicionar um número específico de instâncias ou uma porcentagem do tamanho do grupo existente ou definir o grupo para um tamanho exato.

   Por exemplo, para criar uma política de ajuste de escala em etapas que aumente a capacidade da frota em 30%, escolha `Add`, digite `30` no próximo campo e escolha `percent of group`. Por padrão, o limite inferior desse ajuste em etapas é o limite do alarme, e o limite superior é positivo (+) infinito.
8. Para adicionar outra etapa, escolha Add step (Adicionar etapa) e defina o valor de acordo com o qual dimensionar e os limites inferior e superior da etapa em relação ao limite do alarme.
9. Para definir um número mínimo de instâncias a serem dimensionadas, atualize o campo número em Add capacity units in increments of at least (Adicionar unidades de capacidade em incrementos de pelo menos) 1 capacity units (unidades de capacidade).
10. (Opcional) Em Aquecimento das instâncias, atualize o valor de aquecimento de instância conforme necessário.
11. Escolha Criar.

AWS CLI

Para criar uma política de escala por etapas para aumento a escala horizontalmente (aumento da capacidade), você pode usar os seguintes comandos de exemplo. Substitua cada *`user input placeholder`* por suas próprias informações.

Ao usar o AWS CLI, primeiro você cria uma política de escalabilidade por etapas que fornece instruções ao Amazon EC2 Auto Scaling sobre como escalar quando o valor de uma métrica está aumentando. Nesta etapa, você criará dois alarmes identificando a métrica a ser observada, definindo os limites superior e inferior da métrica e outros detalhes para os alarmes, e associando os alarmes às políticas de dimensionamento.

Etapa 1: Criar uma política para aumentar a escala horizontalmente

Use o put-scaling-policycomando a seguir para criar uma política de escalabilidade de etapas chamadamy-step-scale-out-policy, com um tipo de ajuste PercentChangeInCapacity que aumente a capacidade do grupo com base nos seguintes ajustes de etapa (supondo um limite de CloudWatch alarme de 60%):

- Aumentar a contagem de instâncias em 10 por cento quando o valor da métrica for maior que ou igual a 60 por cento, mas menor que 75 por cento
- Aumentar a contagem de instâncias em 20 por cento quando o valor da métrica for maior que ou igual a 75 por cento, mas menor que 85 por cento
- Aumentar a contagem de instâncias em 30 por cento quando o valor da métrica for maior ou igual 85 por cento

```
aws autoscaling put-scaling-policy \
  --auto-scaling-group-name my-asg  \
  --policy-name my-step-scale-out-policy \
  --policy-type StepScaling \
  --adjustment-type PercentChangeInCapacity \
  --metric-aggregation-type Average \
  --step-adjustments
 MetricIntervalLowerBound=0.0,MetricIntervalUpperBound=15.0,ScalingAdjustment=10 \

 MetricIntervalLowerBound=15.0,MetricIntervalUpperBound=25.0,ScalingAdjustment=20 \
                     MetricIntervalLowerBound=25.0,ScalingAdjustment=30 \
  --min-adjustment-magnitude 1
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa dele para criar um CloudWatch alarme para a política.

```
{
    "PolicyARN":
 "arn:aws:autoscaling:region:123456789012:scalingPolicy:4ee9e543-86b5-4121-b53b-
aa4c23b5bbcc:autoScalingGroupName/my-asg:policyName/my-step-scale-in-policy
}
```

Etapa 2: criar um CloudWatch alarme para o limite máximo métrico

Use o CloudWatch put-metric-alarmcomando a seguir para criar um alarme que aumente o tamanho do grupo de Auto Scaling com base em um valor limite médio de CPU de 60% por pelo

menos dois períodos consecutivos de avaliação de dois minutos. Para usar sua própria métrica personalizada, especifique o nome em `--metric-name` e o namespace em `--namespace`.

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmHigh-AddCapacity \
  --metric-name CPUUtilization --namespace AWS/EC2 --statistic Average \
  --period 120 --evaluation-periods 2 --threshold 60 \
  --comparison-operator GreaterThanOrEqualToThreshold \
  --dimensions "Name=AutoScalingGroupName,Value=my-asg" \
  --alarm-actions PolicyARN
```

Como criar uma política de escalabilidade em etapas para redução

Para criar uma política de escala por etapas para reduzir a escala horizontalmente do seu grupo do Auto Scaling, use um dos métodos a seguir:

Console

Etapa 1: criar um CloudWatch alarme para o limite métrico baixo

1. Abra o CloudWatch console em https://console.aws.amazon.com/cloudwatch/.
2. Se necessário, altere a região da . Na barra de navegação, selecione a região na qual o grupo do Auto Scaling reside.
3. No painel de navegação, escolha Alarms, All alarms (Alarmes, Todos os alarmes) e Create alarm (Criar alarme).
4. Escolha Selecionar métrica.
5. Na guia All metrics (Todas as métricas), escolha EC2, By Auto Scaling Group (Por grupo do Auto Scaling) e insira o nome do grupo do Auto Scaling no campo de pesquisa. Depois, selecione `CPUUtilization` e escolha Selecionar métrica. A página Especificar métrica e condições será exibida, mostrando um gráfico e outras informações sobre a métrica.
6. Em Period (Período), escolha o período de avaliação para o alarme, por exemplo, 1 minuto. Ao avaliar o alarme, todos os períodos são agregados em um único ponto de dados.

   > **Note**
   >
   > Um período mais curto cria um alarme mais sensível.

7. Em Condições, faça o seguinte:

- Em Tipo de limite, escolha Estático.
- Em Whenever **`CPUUtilization`** is (Sempre que for), especifique se você deseja que o valor da métrica seja maior que, maior que ou igual a, menor que, ou menor que ou igual ao limite de violação do alarme. Em than (que), insira o valor do limite desejado de violação de alarme.

> **Important**
>
> Para um alarme a ser usado com uma política para reduzir a escala horizontalmente (alarme inferior), certifique-se de não escolher um valor maior que ou igual ao limite.

8. Em Configuração adicional, faça o seguinte:

   - Em Datapoints to alarm (Pontos de dados para alarme), insira o número de pontos de dados (períodos de avaliação) durante os quais o valor da métrica deverá atender às condições de limite para o alarme. Por exemplo, com dois períodos consecutivos de 5 minutos, o estado de alarme levaria 10 minutos para ser invocado.
   - Em Tratamento de dados ausentes, escolha Tratar dados ausentes como inválidos (limite de violação). Para obter mais informações, consulte Configurando como CloudWatch os alarmes tratam dados perdidos no Guia CloudWatch do usuário da Amazon.

9. Escolha Próximo.

   A página Configure actions (Configurar ações) é exibida.

10. Em Notification (Notificação), selecione um tópico do Amazon SNS para notificar quando o alarme estiver no estado `ALARM`, `OK` ou `INSUFFICIENT_DATA`.

    Para que o alarme envie várias notificações para o mesmo estado de alarme ou para diferentes estados de alarme, escolha Add notification (Adicionar notificação).

    Para que o alarme não envie notificações, escolha Remove (Remover).

11. Você pode deixar vazias as outras seções da página Configure actions (Configurar ações). Deixar as outras seções vazias cria um alarme sem associá-lo a uma política de escalabilidade. Em seguida, você pode associar o alarme a uma política de escalabilidade do console do Amazon EC2 Auto Scaling.

12. Escolha Próximo.

13. Insira um nome (por exemplo, `Step-Scaling-AlarmLow-RemoveCapacity`) e, opcionalmente, uma descrição para o alarme e escolha Próximo.
14. Selecione Criar alarme.

Use o procedimento a seguir para continuar de onde parou depois de criar o CloudWatch alarme.

Etapa 2: Criar uma política de escalabilidade para reduzir a escala horizontalmente

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Verificar se os limites de escalabilidade estão definidos adequadamente. Por exemplo, se sua capacidade desejada já estiver no máximo, especifique um novo máximo para aumentar a escala horizontalmente. Para obter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.
4. Na guia Automatic scaling (Escalabilidade automática), em Dynamic scaling policies (Políticas dinâmicas de escalabilidade), selecione Create dynamic scaling policy (Criar política dinâmica de escalabilidade).
5. Para Tipo de política, escolha Escala por etapas e especifique um nome para a política.
6. Para CloudWatch alarme, escolha seu alarme. Se você ainda não criou um alarme, escolha Criar um CloudWatch alarme e conclua as etapas 4 a 14 no procedimento anterior para criar um alarme.
7. Especifique a alteração no tamanho do grupo atual que essa política fará quando executada usando Take the action (Executar a ação). É possível remover um número específico de instâncias ou uma porcentagem do tamanho do grupo existente ou definir o grupo para um tamanho exato.

   Por exemplo, para criar uma política de redução da escala horizontalmente que diminua a capacidade do grupo em duas instâncias, selecione `Remove`, digite 2 no próximo campo e, em seguida, selecione `capacity units`. Por padrão, o limite superior desse ajuste em etapas é o limite do alarme, e o limite inferior é negativo (-) infinito.
8. Para adicionar outra etapa, escolha Add step (Adicionar etapa) e defina o valor de acordo com o qual dimensionar e os limites inferior e superior da etapa em relação ao limite do alarme.

9. Escolha Criar.

AWS CLI

Para criar uma política de escalonamento de etapas para reduzir a escala horizontalmente (diminuir a capacidade), você pode usar os seguintes comandos de exemplo. Substitua cada *user input placeholder* por suas próprias informações.

Ao usar o AWS CLI, primeiro você cria uma política de escalabilidade por etapas que fornece instruções ao Amazon EC2 Auto Scaling sobre como escalar quando o valor de uma métrica está diminuindo. Nesta etapa, você criará dois alarmes identificando a métrica a ser observada, definindo os limites superior e inferior da métrica e outros detalhes para os alarmes, e associando os alarmes às políticas de dimensionamento.

Etapa 1: Criar uma política para reduzir a escala horizontalmente

Use o put-scaling-policycomando a seguir para criar uma política de escalabilidade de etapas chamadamy-step-scale-in-policy, com um tipo de ajuste ChangeInCapacity que diminui a capacidade do grupo em 2 instâncias quando o CloudWatch alarme associado ultrapassa o valor mínimo do limite métrico.

```
aws autoscaling put-scaling-policy \
  --auto-scaling-group-name my-asg  \
  --policy-name my-step-scale-in-policy \
  --policy-type StepScaling \
  --adjustment-type ChangeInCapacity \
  --step-adjustments MetricIntervalUpperBound=0.0,ScalingAdjustment=-2
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa dele para criar o CloudWatch alarme para a política.

```
{
    "PolicyARN": "arn:aws:autoscaling:region:123456789012:scalingPolicy:ac542982-
cbeb-4294-891c-a5a941dfa787:autoScalingGroupName/my-asg:policyName/my-step-scale-
out-policy
}
```

Etapa 2: criar um CloudWatch alarme para o limite métrico baixo

Use o CloudWatch put-metric-alarmcomando a seguir para criar um alarme que diminua o tamanho do grupo de Auto Scaling com base no valor limite médio de CPU de 40 por cento por pelo menos dois períodos de avaliação consecutivos de dois minutos. Para usar sua própria métrica personalizada, especifique o nome em `--metric-name` e o namespace em `--namespace`.

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmLow-RemoveCapacity \
  --metric-name CPUUtilization --namespace AWS/EC2 --statistic Average \
  --period 120 --evaluation-periods 2 --threshold 40 \
  --comparison-operator LessThanOrEqualToThreshold \
  --dimensions "Name=AutoScalingGroupName,Value=my-asg" \
  --alarm-actions PolicyARN
```

Políticas de escalabilidade simples

Os exemplos a seguir mostram como você pode usar os comandos da CLI para criar políticas de escala simples. Elas permanecem neste documento como referência para os clientes que quiserem usá-las, mas recomendamos que você use políticas de monitoramento de destino ou de escala por etapas.

Semelhantes às políticas de escalabilidade por etapas, as políticas de escalabilidade simples exigem que você crie CloudWatch alarmes para suas políticas de escalabilidade. Ambas exigem que você defina se deseja adicionar ou remover instâncias, e quantas delas, ou defina o grupo para um tamanho exato.

Uma das principais diferenças entre as políticas de escala por etapas e as políticas de escala simples são os ajustes por etapas que você obtém com as políticas de escala por etapas. Com a escala por etapas, é possível fazer alterações maiores ou menores no tamanho do grupo com base nos ajustes de etapas que você especificar.

Uma política de escala simples também deve aguardar a conclusão de uma atividade de escala em andamento ou a substituição da verificação de integridade e o término de um período de espera antes de responder a alarmes adicionais. Em contraste, com a escalabilidade em etapas, a política pode continuar a responder a alarmes adicionais, mesmo enquanto a substituição de uma ação de escalabilidade ou uma verificação de integridade está em andamento. Isso significa que o Amazon EC2 Auto Scaling avalia todas as violações de alarme à medida que recebe as mensagens de alarme. Recomendamos que você use as políticas de escalabilidade em etapas em vez da simples, mesmo que você tenha um único ajuste em etapas.

O Amazon EC2 Auto Scaling originalmente oferecia suporte apenas a políticas de escalabilidade simples. Se você criou sua política de escalabilidade antes de as políticas em etapa e de rastreamento de destino serem introduzidas, sua política será tratada como uma simples política de escalabilidade.

Como criar uma política de escalabilidade em etapas para expansão

Use o put-scaling-policycomando a seguir para criar uma política de escalabilidade simples chamadamy-simple-scale-out-policy, com um tipo de ajuste `PercentChangeInCapacity` que aumenta a capacidade do grupo em 30% quando o CloudWatch alarme associado ultrapassa o valor máximo do limite métrico.

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-out-policy \
  --auto-scaling-group-name my-asg --scaling-adjustment 30 \
  --adjustment-type PercentChangeInCapacity
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa dele para criar o CloudWatch alarme para a política.

Como criar uma política de escalabilidade em etapas para redução

Use o put-scaling-policycomando a seguir para criar uma política de escalabilidade simples chamadamy-simple-scale-in-policy, com um tipo de ajuste `ChangeInCapacity` que diminui a capacidade do grupo em uma instância quando o CloudWatch alarme associado viola o valor mínimo do limite métrico.

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-in-policy \
  --auto-scaling-group-name my-asg --scaling-adjustment -1 \
  --adjustment-type ChangeInCapacity --cooldown 180
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa dele para criar o CloudWatch alarme para a política.

Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling

> **Important**
>
> Como prática recomendada, recomendamos não usar políticas de escalabilidade simples e desaquecimento de escalabilidade. Uma política de escalabilidade com rastreamento do destino ou uma política de escalabilidade em etapas é melhor para a performance da

escalabilidade. Para uma política de escalabilidade que altera o tamanho do grupo do Auto Scaling proporcionalmente à medida que o valor da métrica de escalabilidade diminui ou aumenta, recomendamos o monitoramento do objetivo em escalabilidade simples ou escalabilidade em etapas.

Ao criar políticas de escalabilidade simples para seu grupo do Auto Scaling, recomendamos que você configure o esfriamento de escalabilidade ao mesmo tempo.

Após iniciar ou terminar instâncias, o grupo do Auto Scaling espera o período de desaquecimento encerrar antes que qualquer outra ação de escalabilidade iniciada por políticas de escalabilidade simples possa ser iniciada. A intenção do período de esfriamento é impedir que o grupo do Auto Scaling inicie ou encerre outras instâncias antes que os efeitos de atividades de escalabilidade anteriores sejam visíveis.

Suponha, por exemplo, que uma política de escalabilidade simples para utilização da CPU recomende iniciar duas instâncias. O Amazon EC2 Auto Scaling inicia duas instâncias e pausa as ações de escalabilidade até o período de desaquecimento terminar. Quando o período de desaquecimento terminar, será possível retomar todas as ações de escalabilidade iniciadas por políticas de escalabilidade simples. Se a utilização da CPU violar o limite alto do alarme novamente, o grupo do Auto Scaling aumentará a escala na horizontal novamente, e o período de desaquecimento entrará em vigor novamente. Porém, se duas instâncias forem suficientes para diminuir o valor da métrica, o grupo permanecerá no tamanho atual.

Conteúdo

- Considerações
- Ganchos do ciclo de vida podem causar mais atrasos
- Alterar o período de desaquecimento padrão
- Definir um período de desaquecimento para políticas de escalabilidade simples específicas

Considerações

As considerações a seguir se aplicam ao trabalhar com políticas de escalabilidade simples e desaquecimentos de escalabilidade:

- As políticas de monitoramento do objetivo e escalabilidade de etapas podem iniciar uma ação de aumento da escala na horizontal imediatamente sem esperar que o período de desaquecimento

termine. Em vez disso, sempre que o grupo do Auto Scaling inicia instâncias, as instâncias individuais têm um período de aquecimento. Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

- Enquanto uma atividade de aumento de escala na horizontal estiver em andamento, todas as atividades de rastreamento de destino e etapas de redução da escala na horizontal serão bloqueadas até que as instâncias terminem o aquecimento. As atividades de redução da escala na horizontal também podem ser adiadas quando um grupo do Auto Scaling estiver em um período de espera.
- Quando uma ação programada começa no horário programado, ela pode acionar uma ação de escalabilidade imediatamente, sem esperar que o período de desaquecimento termine.
- Se uma instância se tornar não íntegra, o Amazon EC2 Auto Scaling não aguardará o fim do período de desaquecimento para substituir a instância não íntegra.
- Quando várias instâncias são iniciadas ou terminadas, o período de desaquecimento (o desaquecimento padrão ou o desaquecimento específico da política de escalabilidade) entra em vigor quando a última instância conclui seu início ou término.
- Quando o grupo do Auto Scaling é escalado manualmente, o padrão é não aguardar o desaquecimento terminar. No entanto, você pode ignorar esse comportamento e respeitar o tempo de recarga padrão ao usar o AWS CLI ou um SDK para escalar manualmente.
- Por padrão, o Elastic Load Balancing aguarda 300 segundos para concluir o processo de cancelamento do registro (descarga da conexão). Se o grupo estiver atrás de um balanceador de carga do Elastic Load Balancing, ele aguardará que as instâncias de encerramento cancelem o registro antes de iniciar o período de desaquecimento.

Ganchos do ciclo de vida podem causar mais atrasos

Caso um gancho do ciclo de vida seja invocado, o período de desaquecimento começará após a conclusão da ação do ciclo de vida ou após o período do tempo limite terminar. Por exemplo, considere um grupo do Auto Scaling com um gancho do ciclo de vida para iniciar a instância. Quando a aplicação passa por um aumento na demanda, o grupo executa uma instância para adicionar capacidade. Como há um gancho do ciclo de vida, a instância é colocada em estado de espera, e as ações de escalabilidade causadas por políticas de escalabilidade simples são pausadas. Quando a instância entra no estado `InService`, o período de desaquecimento é iniciado. Quando o período de desaquecimento termina, atividades de políticas de escalabilidade simples são retomadas.

Quando o Elastic Load Balancing estiver habilitado, para fins de redução da escala na horizontal, o período de desaquecimento será iniciado quando a instância de término encerrar a descarga

da conexão (atraso de cancelamento do registro). O período de espera não aguarda o término da drenagem da conexão ou a conclusão da ação do gancho do ciclo de vida. Isso significa que todas as ações de escalabilidade causadas por políticas de escalabilidade simples podem ser retomadas assim que o resultado do evento de redução na escala na horizontal for refletido na capacidade do grupo. Caso contrário, esperar para concluir todas as três atividades (descarga da conexão, gancho do ciclo de vida e período de desaquecimento) aumentaria consideravelmente a quantidade de tempo de que o grupo do Auto Scaling precisa para pausar a escalabilidade.

Alterar o período de desaquecimento padrão

Não é possível definir o desaquecimento padrão quando você inicialmente cria um grupo do Auto Scaling no console do Amazon EC2 Auto Scaling. Por padrão, esse período de desaquecimento é definido para 300 segundos (5 minutos). Se necessário, você poderá atualizar isso depois que o grupo for criado.

Para alterar o período de desaquecimento padrão (console)

Depois de criar o grupo do Auto Scaling, na guia Details (Detalhes), escolha Advanced configurations (Configurações avançadas), Edit (Editar). Em Default cooldown (Desaquecimento padrão), escolha o período que você deseja com base no tempo de inicialização da instância ou em outras necessidades da aplicação.

Para alterar o período de desaquecimento padrão (AWS CLI)

Use os comandos a seguir para alterar o desaquecimento padrão para grupos do Auto Scaling novos ou existentes. Se o desaquecimento padrão não for definido, será usado o valor padrão de 300 segundos.

- create-auto-scaling-group
- update-auto-scaling-group

Para confirmar o valor do tempo de recarga padrão, use o describe-auto-scaling-groupscomando.

Definir um período de desaquecimento para políticas de escalabilidade simples específicas

Por padrão, todas as políticas de escalabilidade simples usam o período de desaquecimento padrão definido para o grupo do Auto Scaling. Para configurar um período de desaquecimento para poíticas de escalabilidade simples específicas, use o parâmetro de desaquecimento opcional ao criar ou

atualizar a política. Quando um período de desaquecimento é especificado para uma política, ele substitui o desaquecimento padrão.

Um uso comum para um período de desaquecimento específico de política de escalabilidade é com uma política de redução da escala na horizontal. Como essa política termina instâncias, o Amazon EC2 Auto Scaling precisa de menos tempo para determinar se deve terminar instâncias adicionais. Encerrar instâncias deve ser uma operação muito mais rápida do que iniciar instâncias. O desaquecimento padrão de 300 segundos é, portanto, muito longo. Nesse caso, um período de desaquecimento específico de política de escalabilidade com um valor inferior para política de redução da escala na horizontal pode ajudar a diminuir custos, permitindo que o grupo reduza a escala na horizontal mais rapidamente.

Para criar ou atualizar políticas de escalabilidade simples no console, escolha a guia Automatic scaling (Escalabilidade automática) depois de criar o grupo. Para criar ou atualizar políticas de escalabilidade simples usando o AWS CLI, use o put-scaling-policycomando. Para obter mais informações, consulte Políticas de escalabilidade simples e em etapas.

Ajuste de escala baseado no Amazon SQS

> **Important**
>
> As informações e etapas a seguir mostram como calcular o backlog de filas do Amazon SQS por instância usando o atributo `ApproximateNumberOfMessages` queue antes de publicá-lo como uma métrica personalizada para. CloudWatch No entanto, é possível economizar o custo e o esforço investidos na publicação de sua própria métrica usando a matemática em métricas. Para obter mais informações, consulte Criação de uma política de escalabilidade com monitoramento de destino com a matemática em métricas.

É possível fazer a escalabilidade do grupo do Auto Scaling em resposta às alterações na carga do sistema em uma fila do Amazon Simple Queue Service (Amazon SQS). Para saber mais sobre como você pode usar o Amazon SQS, consulte o Guia do desenvolvedor do Amazon Simple Queue Service.

Há alguns cenários em que se pode cogitar a escalabilidade em resposta à atividade em uma fila do Amazon SQS. Por exemplo, suponha que você tenha uma aplicação Web que permita aos usuários fazer upload de imagens e usá-las online. Nesse cenário, cada imagem requer redimensionamento e codificação antes de poder ser publicada. A aplicação é executada em instâncias do EC2 em um grupo do Auto Scaling e é configurada para lidar com as taxas típicas de upload. Instâncias

não íntegras são encerradas e substituídas para manter os níveis de instância atuais em todos os momentos. A aplicação coloca os dados de bitmap brutos das imagens em uma fila do SQS para processamento. Ela processa as imagens e, em seguida, publica as imagens processadas onde possam ser visualizadas pelos usuários. A arquitetura desse cenário funcionará bem se o número de uploads de imagem não variar ao longo do tempo. No entanto, se o número de uploads mudar ao longo do tempo, você pode considerar o uso da escalabilidade dinâmica para dimensionar a capacidade do grupo do Auto Scaling.

Conteúdo

- Usar o monitoramento do objetivo com a métrica correta
- Limitações
- Configurar escalabilidade baseada no Amazon SQS
- Amazon SQS e proteção contra redução de escala na horizontal de instâncias

Usar o monitoramento do objetivo com a métrica correta

Se você usar uma política de escalabilidade com monitoramento de objetivo baseada em uma métrica de fila do Amazon SQS personalizada, a escalabilidade dinâmica poderá se ajustar à curva de demanda da aplicação de forma mais eficaz. Para obter mais informações sobre como escolher métricas para rastreamento de destino, consulte Escolher métricas.

O problema com o uso de uma métrica do CloudWatch Amazon SQS, como `ApproximateNumberOfMessagesVisible` para rastreamento de metas, é que o número de mensagens na fila pode não mudar proporcionalmente ao tamanho do grupo de Auto Scaling que processa as mensagens da fila. Isso ocorre porque número de mensagens na fila do SQS não define exclusivamente o número de instâncias necessário. O número de instâncias no grupo do Auto Scaling pode ser determinado por vários fatores, incluindo o tempo necessário para processar uma mensagem e a quantidade de latência (atraso na fila) aceitável.

A solução é usar uma métrica backlog por instância com o valor de destino sendo o backlog aceitável por instância a ser mantido. Você pode calcular esses números da seguinte maneira:

- Backlog por instância: para calcular o backlog por instância, comece com o atributo da fila `ApproximateNumberOfMessages` para determinar o comprimento da fila do SQS (número de mensagens disponíveis para recuperação da fila). Divida esse número pela capacidade de execução da frota, que para um grupo do Auto Scaling é o número de instâncias no estado `InService`, para obter o backlog por instância.

- Backlog aceitável por instância: para calcular o valor de destino, primeiro determine o que a aplicação pode aceitar em termos de latência. Depois, pegue o valor de latência aceitável e divida-o pelo tempo médio que uma instância do EC2 leva para processar uma mensagem.

Como exemplo, digamos que você tenha um grupo do Auto Scaling com 10 instâncias e o número de mensagens visíveis na fila (`ApproximateNumberOfMessages`) seja de 1.500. Se o tempo médio de processamento for de 0,1 segundo para cada mensagem e a latência mais longa aceitável for de 10 segundos, o backlog aceitável por instância será 10/0,1, o que equivale a 100 mensagens. Isso significa que 100 é o valor de destino da sua política de rastreamento de destino. O evento de aumento horizontal da escala ocorrerá quando o backlog por instância atingir o valor desejado. Como o backlog por instância já está em 150 mensagens (1.500 mensagens/10 instâncias), o grupo passa por um aumento da escala na horizontal com 5 instâncias para manter a proporção em relação ao valor do objetivo.

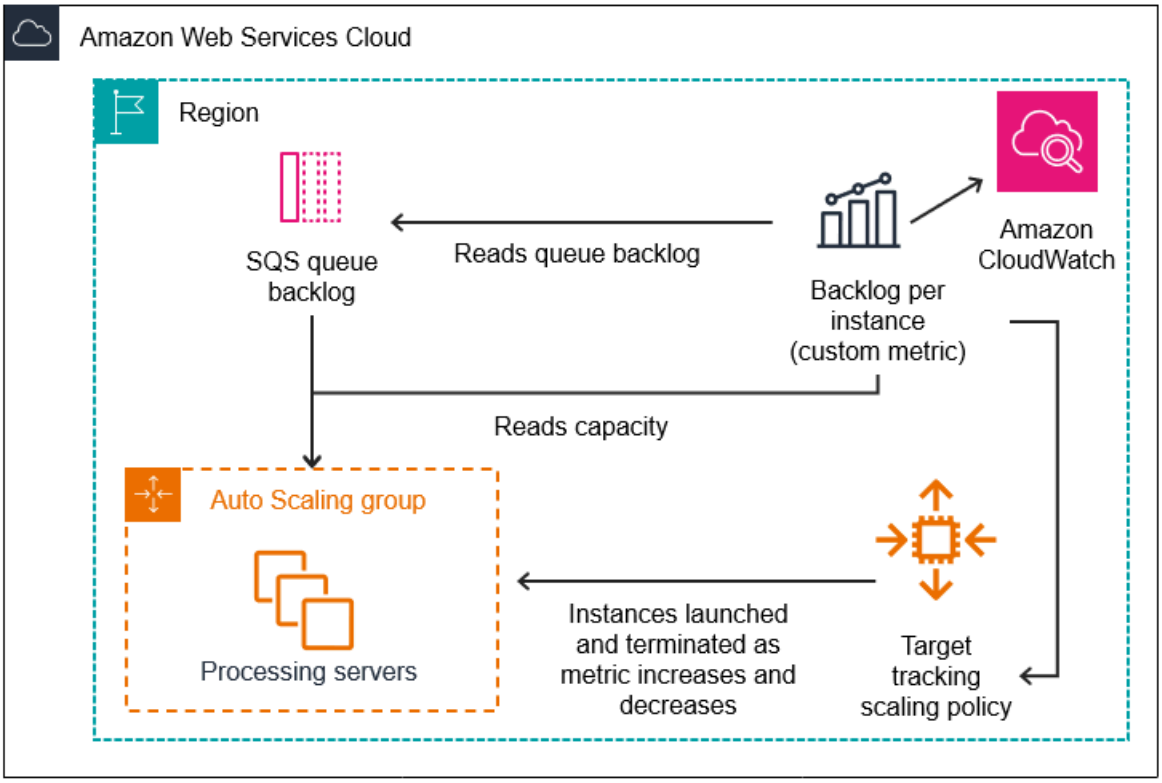
Os procedimentos a seguir demonstram como publicar a métrica personalizada e criar a política de escalabilidade com monitoramento do objetivo que configura o grupo do Auto Scaling para escalar com base nesses cálculos.

> **Important**
>
> Para reduzir custos, lembre-se de usar a matemática em métricas. Para obter mais informações, consulte Criação de uma política de escalabilidade com monitoramento de destino com a matemática em métricas.

Existem três partes principais nessa configuração:

- Um grupo do Auto Scaling para gerenciar instâncias do EC2 para fins de processamento de mensagens de uma fila do SQS.
- Uma métrica personalizada para enviar à Amazon CloudWatch que mede o número de mensagens na fila por instância do EC2 no grupo Auto Scaling.
- Uma política de rastreamento de metas que configura seu grupo de Auto Scaling para escalar com base na métrica personalizada e em um valor alvo definido. CloudWatch os alarmes invocam a política de escalabilidade.

O diagrama a seguir ilustra a arquitetura dessa configuração.



Limitações

Você deve usar o AWS CLI ou um SDK para publicar sua métrica personalizada no. CloudWatch Em seguida, você pode monitorar sua métrica com Console de gerenciamento da AWS o.

Nas seções a seguir, você usa o AWS CLI para as tarefas que precisa realizar. Por exemplo, para obter dados métricos que reflitam o uso atual da fila, você usa o comando SQS. get-queue-attributes

Configurar escalabilidade baseada no Amazon SQS

O procedimento a seguir descreve como configurar a escalabilidade com base no Amazon SQS. Você aprenderá a criar uma métrica CloudWatch personalizada, a configurar uma política de rastreamento de metas usando o AWS CLI e a testar sua configuração.

Antes de começar, verifique se você tem o AWS CLI instalado e configurado. Além disso, é necessário ter uma fila do Amazon SQS para usar. Nas tarefas a seguir, supõe-se que você já tenha uma fila (padrão ou FIFO), um grupo do Auto Scaling e instâncias do EC2 executando a aplicação que usa a fila.

Para obter mais informações sobre o Amazon SQS, consulte o Guia do desenvolvedor do Amazon Simple Queue Service.

Tarefas

- Etapa 1: criar uma métrica CloudWatch personalizada
- Etapa 2: Criar uma política de escalabilidade com monitoramento do objetivo
- Etapa 3: Testar sua política de escalabilidade

Etapa 1: criar uma métrica CloudWatch personalizada

Uma métrica personalizada é definida usando um nome de métrica e um namespace de sua escolha. Namespaces para métricas personalizadas não podem começar com `AWS/`. Para obter mais informações sobre a publicação de métricas personalizadas, consulte o tópico Publicar métricas personalizadas no Guia CloudWatch do usuário da Amazon.

Siga este procedimento para criar a métrica personalizada lendo primeiro as informações da sua AWS conta. Depois, calcule a métrica de backlog por instância, conforme recomendado em uma seção anterior. Por fim, publique esse número com uma CloudWatch granularidade de 1 minuto. Sempre que possível, é altamente recomendável que você escale as métricas com uma granularidade de um minuto para garantir uma resposta mais rápida às alterações na carga do sistema.

Para criar uma métrica CloudWatch personalizada (AWS CLI)

1. Use o comando get-queue-attributes do SQS para obter o número de mensagens em espera na fila (`ApproximateNumberOfMessages`):

   ```
   aws sqs get-queue-attributes --queue-url https://
   sqs.region.amazonaws.com/123456789/MyQueue \
     --attribute-names ApproximateNumberOfMessages
   ```

2. Use o comando describe-auto-scaling-groups para obter a capacidade de execução do grupo, que é o número de instâncias no estado do ciclo de vida `InService`. Esse comando retorna as instâncias de um grupo do Auto Scaling juntamente com seu estado de ciclo de vida.

   ```
   aws autoscaling describe-auto-scaling-groups --auto-scaling-group-names my-asg
   ```

3. Calcule o backlog por instância dividindo o número aproximado de mensagens disponíveis para recuperação da fila pela capacidade de execução do grupo.
4. Crie um script que seja executado a cada minuto para recuperar o valor do backlog por instância e publicá-lo em uma métrica CloudWatch personalizada. Ao publicar uma métrica personalizada,

você especifica o nome da métrica, o espaço nominal, a unidade, o valor e zero ou mais dimensões. Uma dimensão consiste em um nome e um valor de dimensão.

Para publicar sua métrica personalizada, substitua os valores do espaço reservado pelo nome da métrica preferida, pelo valor da métrica, por um namespace (desde que não comece com "AWS“) e pelas dimensões (opcional) e execute o comando a seguir. *italics* put-metric-data

```
aws cloudwatch put-metric-data --metric-name MyBacklogPerInstance --
namespace MyNamespace \
  --unit None --value 20 --
dimensions MyOptionalMetricDimensionName=MyOptionalMetricDimensionValue
```

Depois que seu aplicativo estiver emitindo a métrica desejada, os dados serão enviados para o CloudWatch. A métrica é visível no CloudWatch console. Você pode acessá-lo fazendo login Console de gerenciamento da AWS e navegando até a CloudWatch página. Depois, visualize a métrica navegando até a página de métricas ou procurando-a usando a caixa de pesquisa. Para obter informações sobre métricas de visualização, consulte Visualizar métricas disponíveis no Guia CloudWatch do usuário da Amazon.

Etapa 2: Criar uma política de escalabilidade com monitoramento do objetivo

A métrica que você criou agora pode ser adicionada a uma política de escalabilidade com rastreamento de destino.

Para criar uma política de escalabilidade com rastreamento do destino (AWS CLI)

1. Use o comando `cat` a seguir para especificar um valor de destino para sua política de escalabilidade e uma especificação de métrica personalizada em um arquivo JSON chamado `config.json` em seu diretório inicial. Substitua cada *user input placeholder* por suas próprias informações. Para o `TargetValue`, calcule a métrica backlog aceitável por instância e insira-a aqui. Para calcular esse número, decida um valor de latência normal e divida-o pelo tempo médio necessário para processar uma mensagem, conforme descrito em uma seção anterior.

   Se você não especificou nenhuma dimensão para a métrica criada na etapa 1, não inclua nenhuma dimensão na especificação métrica personalizada.

   ```
   $ cat ~/config.json
   {
   ```

```
   "TargetValue":100,
   "CustomizedMetricSpecification":{
      "MetricName":"MyBacklogPerInstance",
      "Namespace":"MyNamespace",
      "Dimensions":[
         {
            "Name":"MyOptionalMetricDimensionName",
            "Value":"MyOptionalMetricDimensionValue"
         }
      ],
      "Statistic":"Average",
      "Unit":"None"
   }
}
```

2. Use o comando put-scaling-policy, juntamente com o arquivo `config.json` criado na etapa anterior, para criar sua política de escalabilidade.

```
aws autoscaling put-scaling-policy --policy-name sqs100-target-tracking-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://~/config.json
```

   Isso cria dois alarmes: um para escalabilidade e outro para dimensionamento. Ele também retorna o Amazon Resource Name (ARN) da política registrada CloudWatch, que CloudWatch usa para invocar a escalabilidade sempre que o limite métrico é violado.

Etapa 3: Testar sua política de escalabilidade

Depois que a configuração estiver concluída, verifique se a sua política de escalabilidade está funcionando. É possível testá-la aumentando o número de mensagens na fila do SQS e verificando se o grupo do Auto Scaling iniciou uma instância do EC2 adicional. Também é possível testá-la diminuindo o número de mensagens na fila do SQS e verificando se o grupo do Auto Scaling terminou uma instância do EC2.

Para testar a função de expansão

1. Siga as etapas em Criar uma fila padrão do Amazon SQS e enviar uma mensagem ou Criar uma fila FIFO do Amazon SQS e enviar uma mensagem para adicionar mensagens à sua fila.

Certifique-se de que você aumentou o número de mensagens na fila para que a métrica backlog por instância exceda o valor de destino.

Pode levar alguns minutos para que as alterações invoquem o alarme.

2. Use o comando describe-auto-scaling-groups para verificar se o grupo executou uma instância.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Para testar a função de redução

1. Siga as etapas em Receber e excluir mensagens (console) para excluir mensagens da fila. Certifique-se de que você diminuiu o número de mensagens na fila para que a métrica backlog por instância não fique abaixo do valor de destino.

   Pode levar alguns minutos para que as alterações invoquem o alarme.
2. Use o comando describe-auto-scaling-groups para verificar se o grupo encerrou uma instância.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Amazon SQS e proteção contra redução de escala na horizontal de instâncias

As mensagens que não foram processadas no momento em que uma instância foi terminada são devolvidas para a fila do SQS na qual elas podem ser processadas por uma outra instância que ainda esteja em execução. Para aplicações em que tarefas de execução longa são executadas, você pode, opcionalmente, usar a proteção de redução de escala na horizontal de instâncias para ter controle sobre quais trabalhadores de fila são terminados quando o grupo do Auto Scaling sofre redução de escala na horizontal.

O pseudocódigo a seguir mostra uma maneira de proteger processos de trabalho orientados por fila de longa execução contra o término da redução de escala na horizontal.

```
while (true)
{
  SetInstanceProtection(False);
  Work = GetNextWorkUnit();
  SetInstanceProtection(True);
  ProcessWorkUnit(Work);
  SetInstanceProtection(False);
```

```
}
```

Para obter mais informações, consulte Projete suas aplicações para lidar com a terminação de instâncias sem problemas.

Verificar uma ação de escalabilidade para um grupo do Auto Scaling

Na seção Amazon EC2 Auto Scaling do console do Amazon EC2, o Activity history (Histórico de atividades) para um grupo do Auto Scaling permite exibir o status atual de uma ação de escalabilidade que esteja em andamento. Quando a ação de escalabilidade estiver concluída, você poderá ver se ela foi bem-sucedida ou não. Isso é particularmente útil quando você está criando grupos do Auto Scaling ou adicionando condições de escalabilidade a grupos existentes.

Quando você adiciona uma etapa do monitoramento do objetivo, ou política de escalabilidade simples ao grupo do Auto Scaling, o Amazon EC2 Auto Scaling começa imediatamente a avaliar a política em relação à métrica. O alarme da métrica passa para o estado ALARM (ALARME) quando a métrica viola o limite em um determinado número de períodos de avaliação. Isso significa que uma política de escalabilidade pode resultar em uma ação de escalabilidade logo após sua criação. Depois de o Amazon EC2 Auto Scaling alterar a capacidade desejada em resposta a uma política de escalabilidade, é possível verificar a ação de escalabilidade em sua conta. Se deseja receber uma notificação por email do Amazon EC2 Auto Scaling informando sobre uma ação de escalabilidade, siga as instruções em Configurar notificações do Amazon SNS para o Amazon EC2 Auto Scaling.

> **Tip**
>
> No procedimento a seguir, você visualiza as seções Activity history (Histórico de atividades) e Instances (Instâncias) do grupo do Auto Scaling. Em ambas, as colunas nomeadas já deverão ser exibidas. Para exibir colunas ocultas ou alterar o número de linhas exibidas, escolha o ícone de engrenagem no canto superior direito de cada seção para abrir o modal de preferências, atualize as configurações conforme necessário e escolha Confirm (Confirmar).

Para visualizar as ações de escalabilidade do seu grupo do Auto Scaling (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.

2. Na barra de navegação na parte superior da tela, escolha a região onde está o seu grupo do Auto Scaling.
3. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
4. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status (Status) mostra se o seu grupo do Auto Scaling iniciou ou encerrou instâncias com êxito, ou se a ação de escalabilidade ainda está em andamento.
5. (Opcional) Se houver muitas ações de escalabilidade, você poderá escolher o ícone > na borda superior do histórico de atividades para ver a próxima página de ações de escalabilidade.
6. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), a coluna Lifecycle (Ciclo de vida) contém o estado das suas instâncias. Após a instância iniciar e todos os ganchos do ciclo de vida terminarem, seu estado do ciclo de vida mudará para `InService`. A coluna Health status (Status de integridade) mostra o resultado da verificação de integridade da instância do EC2 em sua instância.

Para visualizar as ações de escalabilidade do seu grupo do Auto Scaling (AWS CLI)

Use o seguinte comando describe-scaling-activities:

```
aws autoscaling describe-scaling-activities --auto-scaling-group-name my-asg
```

O seguinte é um exemplo de saída.

As ações de escalabilidade são ordenadas por horário de início. As atividades ainda em andamento são descritas primeiro.

```
{
  "Activities": [
    {
      "ActivityId": "5e3a1f47-2309-415c-bfd8-35aa06300799",
      "AutoScalingGroupName": "my-asg",
      "Description": "Terminating EC2 instance: i-06c4794c2499af1df",
      "Cause": "At 2020-02-11T18:34:10Z a monitor alarm TargetTracking-my-asg-AlarmLow-
b9376cab-18a7-4385-920c-dfa3f7783f82 in state ALARM triggered policy my-target-
tracking-policy changing the desired capacity from 3 to 2.  At 2020-02-11T18:34:31Z
 an instance was taken out of service in response to a difference between desired and
 actual capacity, shrinking the capacity from 3 to 2.  At 2020-02-11T18:34:31Z instance
 i-06c4794c2499af1df was selected for termination.",
```

```
      "StartTime": "2020-02-11T18:34:31.268Z",
      "EndTime": "2020-02-11T18:34:53Z",
      "StatusCode": "Successful",
      "Progress": 100,
      "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-west-2a
\"...}",
      "AutoScalingGroupARN": "arn"
    },
...
  ]
}
```

Para obter uma descrição dos campos na saída, consulte Atividade na Referência da API do Amazon EC2 Auto Scaling.

Para obter ajuda para recuperar as atividades de dimensionamento para um grupo excluído e obter informações sobre os tipos de erros que você pode encontrar e como tratá-los, consulte Solucionar problemas do Amazon EC2 Auto Scaling.

Desabilitar uma política de escalabilidade para um grupo do Auto Scaling

Este tópico descreve como desabilitar temporariamente uma política de escalabilidade para que ela não inicie alterações no número de instâncias no grupo do Auto Scaling. Quando você desabilita uma política de escalabilidade, os detalhes de configuração são preservados, para que seja possível habilitar novamente e rapidamente a política. Isso é mais fácil do que excluir temporariamente uma política quando ela não é necessária e recriá-la mais tarde.

Quando uma política de escalabilidade é desabilitada, o grupo do Auto Scaling não sofre aumento ou redução de escala na horizontal para os alarmes de métrica que são violados enquanto a política de escalabilidade está desabilitada. No entanto, as ações de escalabilidade ainda em andamento não são interrompidas.

Observe que as políticas de escalabilidade desabilitadas ainda são contabilizadas em relação às suas cotas para o número de políticas de escalabilidade que podem ser adicionadas a um grupo do Auto Scaling.

Para desabilitar uma política de escalabilidade (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.

2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Automatic scaling (Escalabilidade automática), em Dynamic scaling policies (Políticas dinâmicas de dimensionamento), marque a caixa de seleção no canto superior direito da política de escalabilidade desejada.
4. Role até o topo da seção Dynamic scaling policies (Políticas dinâmicas de escalabilidade) e selecione Actions (Ações), Disable (Desabilitar).

Quando estiver pronto para habilitar novamente a política de escalabilidade, repita essas etapas e escolha Actions (Ações) e Enable (Habilitar). Depois que você habilitar novamente uma política de escalabilidade, seu grupo do Auto Scaling poderá iniciar imediatamente uma ação de escalabilidade se houver algum alarme no estado ALARM (ALARME).

Como desabilitar uma política de escalabilidade (AWS CLI)

Use o comando put-scaling-policy com a opção `--no-enabled` da seguinte forma. Especifique todas as opções no comando como você as especificaria ao criar a política.

```
aws autoscaling put-scaling-policy --auto-scaling-group-name my-asg \
   --policy-name my-scaling-policy --policy-type TargetTrackingScaling \
   --estimated-instance-warmup 360 \
   --target-tracking-configuration '{ "TargetValue": 70,
 "PredefinedMetricSpecification": { "PredefinedMetricType":
 "ASGAverageCPUUtilization" } }' \
   --no-enabled
```

Como habilitar novamente uma política de escalabilidade (AWS CLI)

Use o comando put-scaling-policy com a opção `--enabled` da seguinte forma. Especifique todas as opções no comando como você as especificaria ao criar a política.

```
aws autoscaling put-scaling-policy --auto-scaling-group-name my-asg \
   --policy-name my-scaling-policy --policy-type TargetTrackingScaling \
   --estimated-instance-warmup 360 \
   --target-tracking-configuration '{ "TargetValue": 70,
 "PredefinedMetricSpecification": { "PredefinedMetricType":
 "ASGAverageCPUUtilization" } }' \
   --enabled
```

Como descrever uma política de escalabilidade (AWS CLI)

Use o comando describe-policies para verificar o status habilitado de uma política de escalabilidade.

```
aws autoscaling describe-policies --auto-scaling-group-name my-asg \
   --policy-names my-scaling-policy
```

O seguinte é um exemplo de saída.

```
{
    "ScalingPolicies": [
        {
            "AutoScalingGroupName": "my-asg",
            "PolicyName": "my-scaling-policy",
            "PolicyARN": "arn:aws:autoscaling:us-
west-2:123456789012:scalingPolicy:1d52783a-b03b-4710-
bb0e-549fd64378cc:autoScalingGroupName/my-asg:policyName/my-scaling-policy",
            "PolicyType": "TargetTrackingScaling",
            "StepAdjustments": [],
            "Alarms": [
                {
                    "AlarmName": "TargetTracking-my-asg-
AlarmHigh-9ca53fdd-7cf5-4223-938a-ae1199204502",
                    "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-9ca53fdd-7cf5-4223-938a-
ae1199204502"
                },
                {
                    "AlarmName": "TargetTracking-my-asg-AlarmLow-7010c83d-d55a-4a7a-
abe0-1cf8b9de6d6c",
                    "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-7010c83d-d55a-4a7a-
abe0-1cf8b9de6d6c"
                }
            ],
            "TargetTrackingConfiguration": {
                "PredefinedMetricSpecification": {
                    "PredefinedMetricType": "ASGAverageCPUUtilization"
                },
                "TargetValue": 70.0,
                "DisableScaleIn": false
            },
            "Enabled": true
```

```
        }
    ]
}
```

Desabilitar uma política de escalabilidade para um grupo do Auto Scaling

Quando você não precisar mais de uma política de escalabilidade, poderá excluí-la. Dependendo do tipo de política de escalabilidade, talvez você também precise excluir os CloudWatch alarmes. A exclusão de uma política de escalabilidade de rastreamento de metas também exclui todos os alarmes associados. CloudWatch A exclusão de uma política de escalonamento de etapas ou de uma política de escalabilidade simples exclui a ação de alarme subjacente, mas não exclui o CloudWatch alarme, mesmo que não tenha mais uma ação associada.

Para excluir uma política de escalabilidade (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Automatic scaling (Escalabilidade automática), em Dynamic scaling policies (Políticas dinâmicas de dimensionamento), marque a caixa de seleção no canto superior direito da política de escalabilidade desejada.
4. Role até o topo da seção Dynamic scaling policies (Políticas dinâmicas de escalabilidade) e selecione Actions (Ações), Delete (Excluir).
5. Quando a confirmação for solicitada, escolha Sim, excluir.
6. (Opcional) Se você excluiu uma política de escalabilidade por etapas ou uma política de escalabilidade simples, faça o seguinte para excluir o CloudWatch alarme associado à política. É possível ignorar essas subetapas para manter o alarme para uso futuro.

   a. Abra o CloudWatch console em https://console.aws.amazon.com/cloudwatch/.
   b. No painel de navegação, escolha Alarms (Alarmes).
   c. Escolha o alarme (por exemplo, `Step-Scaling-AlarmHigh-AddCapacity`) e escolha Action (Ação) e Delete (Excluir).
   d. Quando a confirmação for solicitada, escolha Excluir.

Para obter as políticas de escalabilidade para um grupo do Auto Scaling (AWS CLI)

Antes de excluir uma política de escalabilidade, use o seguinte comando describe-policies para ver quais políticas de escalabilidade foram criadas para o grupo do Auto Scaling. É possível usar a saída ao excluir a política e os alarmes do CloudWatch.

```
aws autoscaling describe-policies --auto-scaling-group-name my-asg
```

É possível filtrar os resultados pelo tipo de política de escalabilidade usando o parâmetro `--query`. Esta sintaxe para `query` funciona no Linux ou no macOS. No Windows, altere as aspas simples para aspas duplas.

```
aws autoscaling describe-policies --auto-scaling-group-name my-asg
  --query 'ScalingPolicies[?PolicyType==`TargetTrackingScaling`]'
```

O seguinte é um exemplo de saída.

```
[
    {
        "AutoScalingGroupName": "my-asg",
        "PolicyName": "cpu50-target-tracking-scaling-policy",
        "PolicyARN": "PolicyARN",
        "PolicyType": "TargetTrackingScaling",
        "StepAdjustments": [],
        "Alarms": [
            {
                "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e",
                "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
            },
            {
                "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
                "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
            }
        ],
        "TargetTrackingConfiguration": {
            "PredefinedMetricSpecification": {
                "PredefinedMetricType": "ASGAverageCPUUtilization"
            },
```

```
            "TargetValue": 50.0,
            "DisableScaleIn": false
        },
        "Enabled": true
    }
]
```

Para excluir a política de dimensionamento (AWS CLI)

Use o comando delete-scaling-policy.

```
aws autoscaling delete-policy --auto-scaling-group-name my-asg \
  --policy-name cpu50-target-tracking-scaling-policy
```

Para excluir seu CloudWatch alarme (AWS CLI)

Para políticas de escalabilidade escalonadas e simples, use o comando delete-alarms para excluir o CloudWatch alarme associado à política. Você pode ignorar essa etapa para manter o alarme para uso futuro. É possível excluir um ou mais alarmes por vez. Por exemplo, use o comando a seguir para excluir os alarmes `Step-Scaling-AlarmHigh-AddCapacity` e `Step-Scaling-AlarmLow-RemoveCapacity`.

```
aws cloudwatch delete-alarms --alarm-name Step-Scaling-AlarmHigh-AddCapacity Step-Scaling-AlarmLow-RemoveCapacity
```

Exemplo de políticas de escalabilidade para o AWS CLI

Você pode criar políticas de escalabilidade para o Amazon EC2 Auto Scaling por meio Console de gerenciamento da AWS do AWS Command Line Interface ,AWS CLI() ou. SDKs

Os exemplos a seguir mostram como você pode criar políticas de escalabilidade para o Amazon EC2 Auto Scaling com o comando. AWS CLI put-scaling-policy Substitua cada *`user input placeholder`* por suas próprias informações.

Para começar a escrever políticas de escalabilidade usando o AWS CLI, consulte os exercícios introdutórios em e. Políticas de escalabilidade de rastreamento de destino Políticas de escalabilidade simples e em etapas

Exemplo 1: como aplicar uma política de escalabilidade com monitoramento do objetivo com uma especificação de métrica predefinida

```
aws autoscaling put-scaling-policy --policy-name cpu50-target-tracking-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
{
  "TargetValue": 50.0,
  "PredefinedMetricSpecification": {
    "PredefinedMetricType": "ASGAverageCPUUtilization"
  }
}
```

Para obter mais informações, consulte a PredefinedMetricSpecificationReferência da API Amazon EC2 Auto Scaling.

> **Note**
>
> Se o arquivo não estiver no diretório atual, digite o caminho completo para o arquivo. Para obter mais informações sobre a leitura de valores de AWS CLI parâmetros de um arquivo, consulte Carregamento de AWS CLI parâmetros de um arquivo no Guia AWS Command Line Interface do usuário.

Exemplo 2: como aplicar uma política de escalabilidade com monitoramento do objetivo com uma especificação de métrica personalizada

```
aws autoscaling put-scaling-policy --policy-name sqs100-target-tracking-scaling-policy
 \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
{
  "TargetValue": 100.0,
  "CustomizedMetricSpecification": {
    "MetricName": "MyBacklogPerInstance",
    "Namespace": "MyNamespace",
    "Dimensions": [{
      "Name": "MyOptionalMetricDimensionName",
      "Value": "MyOptionalMetricDimensionValue"
    }],
    "Statistic": "Average",
    "Unit": "None"
  }
}
```

Para obter mais informações, consulte a CustomizedMetricSpecificationReferência da API Amazon EC2 Auto Scaling.

Exemplo 3: como aplicar uma política de escalabilidade com monitoramento do objetivo somente para expansão

```
aws autoscaling put-scaling-policy --policy-name alb1000-target-tracking-scaling-policy
 \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
{
  "TargetValue": 1000.0,
  "PredefinedMetricSpecification": {
    "PredefinedMetricType": "ALBRequestCountPerTarget",
    "ResourceLabel": "app/my-alb/778d41231b141a0f/targetgroup/my-alb-target-
group/943f017f100becff"
  },
  "DisableScaleIn": true
}
```

Exemplo 4: como aplicar uma política de escalabilidade em etapas para expansão

```
aws autoscaling put-scaling-policy \
  --auto-scaling-group-name my-asg  \
  --policy-name my-step-scale-out-policy \
  --policy-type StepScaling \
  --adjustment-type PercentChangeInCapacity \
  --metric-aggregation-type Average \
  --step-adjustments
 MetricIntervalLowerBound=10.0,MetricIntervalUpperBound=20.0,ScalingAdjustment=10 \

 MetricIntervalLowerBound=20.0,MetricIntervalUpperBound=30.0,ScalingAdjustment=20 \
                     MetricIntervalLowerBound=30.0,ScalingAdjustment=30 \
  --min-adjustment-magnitude 1
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa do ARN ao criar o CloudWatch alarme.

Exemplo 5: como aplicar uma política de escalabilidade em etapas para redução

```
aws autoscaling put-scaling-policy \
  --auto-scaling-group-name my-asg  \
```

```
  --policy-name my-step-scale-in-policy \
  --policy-type StepScaling \
  --adjustment-type ChangeInCapacity \
  --step-adjustments MetricIntervalUpperBound=0.0,ScalingAdjustment=-2
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa do ARN ao criar o CloudWatch alarme.

Exemplo 6: como aplicar uma política de escalabilidade simples para expansão

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-out-policy \
  --auto-scaling-group-name my-asg --scaling-adjustment 30 \
  --adjustment-type PercentChangeInCapacity --min-adjustment-magnitude 2
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa do ARN ao criar o CloudWatch alarme.

Exemplo 7: como aplicar uma política de escalabilidade simples para redução

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-in-policy \
  --auto-scaling-group-name my-asg --scaling-adjustment -1 \
  --adjustment-type ChangeInCapacity --cooldown 180
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa do ARN ao criar o CloudWatch alarme.

Escala preditiva para o Amazon EC2 Auto Scaling

A escala preditiva funciona por meio da análise de dados históricos de carga para detectar padrões diários ou semanais nos fluxos de tráfego. Ele usa essas informações para prever as necessidades futuras de capacidade, de modo que o Amazon EC2 Auto Scaling possa aumentar proativamente a capacidade do seu grupo do Auto Scaling para combinar com a carga prevista.

A escalabilidade preditiva é adequada para situações em que há:

- Tráfego cíclico, como alta utilização de recursos durante o horário comercial e baixa utilização de recursos durante a noite e nos fins de semana
- Padrões on-and-off de carga de trabalho recorrentes, como processamento em lote, testes ou análise periódica de dados

- Aplicações que demoram muito para inicializar, causando um impacto de latência considerável na performance da aplicação durante eventos de aumento da escala na horizontal

Em geral, se houver padrões regulares de aumento de tráfego e aplicações que demoram muito para inicializar, considere o uso da escalabilidade preditiva. A escalabilidade preditiva pode ajudar você a expandir mais rapidamente, lançando a capacidade antes da carga prevista, em comparação com o uso apenas da escalabilidade dinâmica, que é reativa por natureza. A escalabilidade preditiva também pode economizar dinheiro em sua fatura do EC2 ao ajudar você a evitar a necessidade de provisionar a capacidade de forma excessiva.

Por exemplo, considere uma aplicação com elevado índice de utilização durante o horário comercial e baixo uso durante a noite. No início de cada dia útil, a escalabilidade preditiva pode adicionar capacidade antes do primeiro fluxo de tráfego. Isso ajuda sua aplicação a manter alta disponibilidade e performance ao passar de um período de menor utilização para um período de maior utilização. Você não precisa esperar que a escalabilidade dinâmica reaja à mudança de tráfego. Você também não precisa gastar tempo revisando os padrões de carga da aplicação e tentando alocar a quantidade certa de capacidade usando a escalabilidade programada.

Tópicos

- Como a escalabilidade preditiva funciona
- Desabilitar uma política de escalabilidade para um grupo do Auto Scaling
- Avaliar as políticas de escalabilidade preditiva
- Substituir valores de previsão usando ações programadas
- Configurações avançadas de política de escalabilidade preditiva usando métricas personalizadas

Como a escalabilidade preditiva funciona

Este tópico explica como funciona a escala preditiva e descreve o que deve ser considerado quando você cria uma política de escala preditiva.

Tópicos

- Como funciona
- Limites máximos de capacidade
- Considerações
- Regiões aceitas

Como funciona

Para usar a escala preditiva, crie uma política de escalabilidade preditiva que especifique a CloudWatch métrica a ser monitorada e analisada. Para que a escala preditiva comece a prever valores futuros, essa métrica deve ter pelo menos 24 horas de dados.

Depois que você cria a política, a escala preditiva começa a analisar os dados de métricas dos últimos 14 dias para identificar padrões. Ele usa essa análise para gerar uma previsão horária dos requisitos de capacidade para as próximas 48 horas. A previsão é atualizada a cada 6 horas usando os CloudWatch dados mais recentes. À medida que novos dados chegam, a escala preditiva é capaz de melhorar continuamente a precisão das previsões futuras.

Quando você ativa a escala preditiva pela primeira vez, ela é executada no modo somente de previsão. Nesse modo, ele gera previsões de capacidade, mas não escala de fato seu grupo do Auto Scaling com base nessas previsões. Isso permite que você avalie a exatidão e a adequação da previsão. Você pode visualizar os dados de previsão usando a operação da API `GetPredictiveScalingForecast` ou o Console de gerenciamento da AWS.

Depois de revisar os dados de previsão e decidir iniciar a escala com base nesses dados, alterne a política de escala para o modo de previsão e escala. Neste modo:

- Se a previsão for de um aumento na carga, o Amazon EC2 Auto Scaling aumentará a capacidade por meio do aumento a escala horizontalmente.
- Se a previsão esperar uma redução na carga, ela não será reduzida a escala horizontalmente para remover a capacidade. Se quiser remover a capacidade que não é mais necessária, será necessário criar políticas de escala dinâmica.

Por padrão, o Amazon EC2 Auto Scaling escala seu grupo do Auto Scaling no início de cada hora com base na previsão para essa hora. Opcionalmente, você pode especificar uma hora de início anterior usando a propriedade `SchedulingBufferTime` na operação da API `PutScalingPolicy` ou a configuração de instâncias de pré-execução no Console de gerenciamento da AWS. Isso faz com que o Amazon EC2 Auto Scaling execute novas instâncias antes da demanda prevista, dando-lhes tempo para inicializar e se preparar para lidar com o tráfego.

Para dar suporte à execução de novas instâncias antes da demanda prevista, recomendamos enfaticamente que você ative o aquecimento de instância padrão para o seu grupo do Auto Scaling. Especifica um período de tempo após uma atividade de aumento a escala horizontalmente durante o qual o Amazon EC2 Auto Scaling não reduzirá a escala, mesmo que as políticas de escala dinâmica

indiquem que a capacidade deve ser reduzida. Isso ajuda você a garantir que as instâncias recém-executadas tenham tempo suficiente para começar a atender ao aumento do tráfego antes de serem consideradas para operações de redução da escala horizontalmente. Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Limites máximos de capacidade

Os grupos do Auto Scaling têm uma configuração de capacidade máxima que limita o número máximo de instâncias do EC2 que podem ser executadas para o grupo. Quando as políticas de escalabilidade estão definidas, um grupo não pode aumentar sua capacidade desejada acima do limite da capacidade máxima.

Como alternativa, você pode permitir que a capacidade máxima do grupo seja aumentada automaticamente se a capacidade prevista se aproximar ou exceder a capacidade máxima do grupo do Auto Scaling. Para ativar esse comportamento, use as propriedades `MaxCapacityBreachBehavior` e `MaxCapacityBuffer` na operação da API `PutScalingPolicy` ou a configuração de comportamento de capacidade máxima no Console de gerenciamento da AWS.

> **Warning**
>
> Tenha cuidado ao permitir que a capacidade máxima seja aumentada automaticamente. Isso pode fazer com que mais instâncias sejam executadas do que o pretendido se a capacidade máxima aumentada não for monitorada e gerenciada. A capacidade máxima aumentada passa a ser a nova capacidade máxima normal do grupo do Auto Scaling até que você a atualize manualmente. A capacidade máxima não diminui automaticamente de volta ao máximo original.

Considerações

- Confirme se a escalabilidade preditiva é adequada para sua workload. Uma workload será uma boa opção para o uso da escalabilidade preditiva se ela apresentar padrões de carga recorrentes específicos do dia da semana ou da hora do dia. Para verificar isso, configure políticas de escalabilidade preditiva no modo somente previsão e consulte as recomendações do console. O Amazon EC2 Auto Scaling fornece recomendações com base em observações sobre a performance potencial da política. Avalie a previsão e as recomendações antes de permitir que a escalabilidade preditiva escale ativamente sua aplicação.

- A escalabilidade preditiva precisa de pelo menos 24 horas de dados históricos para começar a previsão. No entanto, as previsões serão mais eficazes se os dados históricos abrangerem duas semanas completas. Se você atualizar sua aplicação criando um novo do grupo do Auto Scaling e excluindo o antigo, o novo grupo do Auto Scaling precisará de 24 horas de dados históricos de carga antes que a escalabilidade preditiva possa começar a gerar previsões novamente. É possível usar métricas personalizadas para agregar métricas em grupos do Auto Scaling novos e antigos. Senão, talvez seja necessário esperar alguns dias para obter uma previsão mais precisa.
- Escolha uma métrica de carga que represente com precisão a carga total da sua aplicação e que seja o aspecto mais importante a ser escalado.
- Usá-la com a escalabilidade preditiva ajuda você a seguir de perto a curva de demanda para a aplicação, reduzindo a escala na horizontal durante períodos de baixo tráfego e aumentando-a quando o tráfego é maior do que o esperado. Quando várias políticas de escalabilidade estão ativas, cada política determina a capacidade desejada de forma independente e a capacidade desejada é definida como a capacidade máxima entre essas. Por exemplo, se 10 instâncias forem necessárias para permanecer na utilização-alvo em uma política de escalabilidade com monitoramento do objetivo e 8 instâncias forem necessárias para permanecer na utilização-alvo em uma política de dimensionamento preditiva, a capacidade desejada do grupo será definida como 10. Também recomendamos usar políticas de ajuste de escala de rastreamento de metas em seu ajuste de escala dinâmico. Para obter mais informações, consulte Escalabilidade dinâmica para o Amazon EC2 Auto Scaling.
- Um pressuposto básico da escalabilidade preditivo é que o grupo do Auto Scaling é homogêneo e todas as instâncias têm capacidade igual. Se isso não for verdade para seu grupo, a capacidade prevista pode ser imprecisa. Portanto, tenha cuidado ao criar políticas de escalabilidade preditiva para grupos de instâncias mistas, já que diferentes tipos de instâncias podem ser provisionados com capacidades diferentes. Veja a seguir alguns exemplos para os quais a capacidade prevista será imprecisa:
  - Sua política de escalabilidade preditiva é baseada na utilização da CPU, mas o número de v CPUs em cada instância do Auto Scaling varia entre os tipos de instância.
  - Sua política de escalabilidade preditiva é baseada na entrada ou na saída da rede, mas throughput de largura de banda da rede para cada instância do Auto Scaling varia entre os tipos de instância. Por exemplo, os tipos de instância M5 e M5n são semelhantes, mas o tipo de instância M5n oferece throughput de rede significativamente maior.

Regiões aceitas

- Leste dos EUA (Norte da Virgínia)
- Leste dos EUA (Ohio)
- Oeste dos EUA (N. da Califórnia)
- Oeste dos EUA (Oregon)
- África (Cidade do Cabo)
- Ásia-Pacífico (Hong Kong)
- Ásia-Pacífico (Hyderabad)
- Ásia-Pacífico (Jacarta)
- Ásia-Pacífico (Melbourne)
- Ásia-Pacífico (Mumbai)
- Ásia-Pacífico (Osaka)
- Ásia-Pacífico (Seul)
- Ásia-Pacífico (Singapura)
- Ásia-Pacífico (Sydney)
- Ásia-Pacífico (Tóquio)
- Canadá (Central)
- Oeste do Canadá (Calgary)
- China (Pequim)
- China (Ningxia)
- Europa (Frankfurt)
- Europa (Irlanda)
- Europa (Londres)
- Europa (Milão)
- Europe (Paris)
- Europa (Espanha)
- Europa (Estocolmo)
- Europa (Zurique)
- Israel (Tel Aviv)
- Oriente Médio (Barém)

- Oriente Médio (Emirados Árabes Unidos)
- América do Sul (São Paulo)
- AWS GovCloud (Leste dos EUA)
- AWS GovCloud (Oeste dos EUA)

Desabilitar uma política de escalabilidade para um grupo do Auto Scaling

Os procedimentos a seguir ajudam você a criar uma política de escalabilidade preditiva usando o Console de gerenciamento da AWS ou. AWS CLI

Se o grupo do Auto Scaling for novo, ele deverá fornecer pelo menos 24 horas de dados antes que o Amazon EC2 Auto Scaling possa gerar uma previsão para ele.

Conteúdo

- Criar uma política de escalabilidade preditiva (console)
- Criar uma política de escalabilidade preditiva (AWS CLI)

Criar uma política de escalabilidade preditiva (console)

Se esta for a primeira vez que cria uma política de escala preditiva, recomendamos usar o console para criar várias políticas de escala preditiva no modo somente de previsão. Isso testa os efeitos potenciais de diferentes métricas e valores de destino. Você pode criar várias políticas de escalabilidade preditiva para cada grupo do Auto Scaling, mas somente uma das políticas pode ser usada para a escalabilidade ativa.

Criar uma política de escalação preditiva no console (métricas predefinidas)

Siga o procedimento a seguir para criar uma política de escalação preditiva usando métricas predefinidas (CPU, E/S da rede ou contagem de solicitações do Application Load Balancer por destino). A maneira mais fácil de criar uma política de escalação preditiva é usar métricas predefinidas. Se você preferir usar métricas personalizadas, consulte Criar uma política de escalação preditiva no console (métricas personalizadas).

Para criar uma política de escalabilidade preditiva

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.

2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.

3. Na guia Automatic scaling (Escalabilidade automática), em Scaling policies (Políticas de escalabilidade), escolha Create predictive scaling policy (Criar política de escalabilidade preditiva).

4. Insira um nome para a política.

5. Ativar a opção Scale based on forecast (Escala baseada em previsão) para conceder ao Amazon EC2 Auto Scaling permissão para começar a escalar imediatamente.

   Para manter a política no modo somente previsão, deixe a opção Scale based on forecast(Escala baseada em previsão) desativada.

6. Em Metrics (Métricas), escolha suas métricas na lista de opções. As opções incluem CPU, Network In (Entrada de rede), Network Out (Saída de rede), Application Load Balancer request count (Número de solicitações do Application Load Balancer) e Custom metric pair (Par de métricas personalizadas).

   Se tiver escolhido Application Load Balancer request count per target (Número de solicitações do Application Load Balancer por destino), escolha um grupo de destino em Target group (Grupo de destino). A opção Application Load Balancer request count per target (Número de solicitações do Application Load Balancer por destino) só será válida de você tiver anexado um grupo de destino do Application Load Balancer ao seu grupo do Auto Scaling.

   Se você escolheu Custom metric pair (Par de métricas personalizadas), escolha métricas individuais nas listas suspensas para Load metric (Métrica de carga) e Scaling metric (Métrica de escalabilidade).

7. Em Target utilization (Utilização-alvo), insira o valor-alvo que o Amazon EC2 Auto Scaling deveria manter. O Amazon EC2 Auto Scaling aumentará a escala na horizontal até que a utilização média seja igual à utilização-alvo ou até atingir o número máximo de instâncias especificado.

Se sua métrica de escalabilidade for...	Então a utilização-alvo representará...
CPU	A porcentagem de CPU que cada instância deve idealmente usar.

Se sua métrica de escalabilidade for...	Então a utilização-alvo representará...
Entrada de rede	O número médio de bytes por minuto que cada instância deve idealmente receber.
Saída de rede	O número médio de bytes por minuto que cada instância deve idealmente enviar.
Número de solicitações do Application Load Balancer por destino	O número médio de solicitações por minuto que cada instância deve idealmente receber.

8. (Opcional) Em Pre-launch instances (Iniciar instâncias previamente), escolha com que antecedência você deseja que suas instâncias sejam iniciadas antes que a previsão solicite o aumento de carga.
9. (Opcional) Em Max capacity behavior (Comportamento na capacidade máxima), escolha se será permitido que o Amazon EC2 Auto Scaling aumente a escala horizontalmente além da capacidade máxima do grupo quando a capacidade prevista exceder o máximo definido. A ativação dessa configuração permite aumentar a escala horizontalmente nos períodos em que estão previstos picos de tráfego.
10. (Opcional) Em Buffer maximum capacity above the forecasted capacity (Capacidade máxima do buffer acima da capacidade prevista), escolha a quantidade de capacidade adicional a ser usada quando a capacidade prevista estiver próxima de ou exceder a capacidade máxima. O valor é especificado como um percentual em relação à capacidade de prevista. Por exemplo, se o buffer é 10, isso significa um buffer de 10%. Portanto, se a capacidade prevista for 50 e a capacidade máxima for 40, a capacidade máxima real será 55.

    Se esse valor for definido como 0, o Amazon EC2 Auto Scaling poderá escalar a capacidade acima da capacidade máxima para igualar, mas não exceder, a capacidade prevista.
11. Selecione Create predictive scaling policy (Criar política de escalabilidade preditiva).

Criar uma política de escalação preditiva no console (métricas personalizadas)

Use o procedimento a seguir para criar uma política de escalação preditiva usando métricas personalizadas. As métricas personalizadas podem incluir outras métricas fornecidas por CloudWatch ou nas quais você publica CloudWatch. Para usar a contagem de solicitações de CPU,

E/S de rede ou Application Load Balancer por destino, consulte Criar uma política de escalação preditiva no console (métricas predefinidas).

Para criar uma política de escalação preditiva usando métricas personalizadas, é necessário fazer o seguinte:

- Você deve fornecer as consultas brutas que permitem que o Amazon EC2 Auto Scaling interaja com as métricas inseridas. CloudWatch Para obter mais informações, consulte Configurações avançadas de política de escalabilidade preditiva usando métricas personalizadas. Para ter certeza de que o Amazon EC2 Auto Scaling pode extrair os CloudWatch dados métricos, confirme se cada consulta está retornando pontos de dados. Confirme isso usando o CloudWatch console ou a operação CloudWatch GetMetricDatada API.

  > **Note**
  >
  > Fornecemos exemplos de cargas úteis JSON no editor de JSON no console do Amazon EC2 Auto Scaling. Esses exemplos fornecem uma referência para os pares de valores-chave necessários para adicionar outras CloudWatch métricas fornecidas por AWS ou nas quais você publicou anteriormente. CloudWatch Você pode usá-las como ponto de partida e depois personalizá-las de acordo com as suas necessidades.

- Se você usar qualquer matemática de métricas, deverá estruturar manualmente o JSON para adequá-lo ao seu cenário específico. Para obter mais informações, consulte Usar expressões de matemática métrica. Antes de usar matemática de métricas em sua política, confirme se as consultas de métricas baseadas em expressões matemáticas de métricas são válidas e retornam uma única série temporal. Confirme isso usando o CloudWatch console ou a operação CloudWatch GetMetricDatada API.

Se você cometer um erro em uma consulta fornecendo dados incorretos, como o nome errado do grupo do Auto Scaling, a previsão não terá nenhum dado. Para solucionar problemas de métricas personalizadas, consulte Criar uma política de escalação preditiva no console (métricas personalizadas).

Para criar uma política de escalabilidade preditiva

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

Um painel dividido é aberto na parte inferior da página.

3. Na guia Automatic scaling (Escalabilidade automática), em Scaling policies (Políticas de escalabilidade), escolha Create predictive scaling policy (Criar política de escalabilidade preditiva).
4. Insira um nome para a política.
5. Ativar a opção Scale based on forecast (Escala baseada em previsão) para conceder ao Amazon EC2 Auto Scaling permissão para começar a escalar imediatamente.

   Para manter a política no modo somente previsão, deixe a opção Scale based on forecast(Escala baseada em previsão) desativada.
6. Em Metrics (Métricas), escolha Custom metric pair (Par de métricas personalizado).

   a. Em Métrica de carga, escolha CloudWatch Métrica personalizada para usar uma métrica personalizada. Estruture a carga útil JSON que contém a definição da métrica de carga para a política e cole-a na caixa do editor de Jason, substituindo o que já está na caixa.

   b. Em Métrica de escala, escolha CloudWatch Métrica personalizada para usar uma métrica personalizada. Estruture a carga útil JSON que contém a definição da métrica de escalação para a política e cole-a na caixa do editor de Jason, substituindo o que já está na caixa.

   c. (Opcional) Para adicionar uma métrica de capacidade personalizada, marque a caixa de seleção Add custom capacity metric (Adicionar métrica de capacidade personalizada). Estruture a carga útil JSON que contém a definição da métrica de capacidade para a política e cole-a na caixa do editor de Jason, substituindo o que já está na caixa.

      Você só precisa habilitar essa opção para criar uma nova série temporal de capacidade se seus dados métricos de capacidade abrangerem vários grupos do Auto Scaling. Nesse caso, é necessário usar a matemática de métricas para agregar os dados em uma única série temporal.
7. Em Target utilization (Utilização-alvo), insira o valor-alvo que o Amazon EC2 Auto Scaling deveria manter. O Amazon EC2 Auto Scaling aumentará a escala na horizontal até que a utilização média seja igual à utilização-alvo ou até atingir o número máximo de instâncias especificado.
8. (Opcional) Em Pre-launch instances (Iniciar instâncias previamente), escolha com que antecedência você deseja que suas instâncias sejam iniciadas antes que a previsão solicite o aumento de carga.

9. (Opcional) Em Max capacity behavior (Comportamento na capacidade máxima), escolha se será permitido que o Amazon EC2 Auto Scaling aumente a escala horizontalmente além da capacidade máxima do grupo quando a capacidade prevista exceder o máximo definido. A ativação dessa configuração permite aumentar a escala horizontalmente nos períodos em que estão previstos picos de tráfego.
10. (Opcional) Em Buffer maximum capacity above the forecasted capacity (Capacidade máxima do buffer acima da capacidade prevista), escolha a quantidade de capacidade adicional a ser usada quando a capacidade prevista estiver próxima de ou exceder a capacidade máxima. O valor é especificado como um percentual em relação à capacidade de prevista. Por exemplo, se o buffer é 10, isso significa um buffer de 10%. Portanto, se a capacidade prevista for 50 e a capacidade máxima for 40, a capacidade máxima real será 55.

    Se esse valor for definido como 0, o Amazon EC2 Auto Scaling poderá escalar a capacidade acima da capacidade máxima para igualar, mas não exceder, a capacidade prevista.
11. Selecione Create predictive scaling policy (Criar política de escalabilidade preditiva).

Criar uma política de escalabilidade preditiva (AWS CLI)

Use o AWS CLI seguinte para configurar políticas de escalabilidade preditiva para seu grupo de Auto Scaling. Substitua cada *user input placeholder* por suas próprias informações.

Para obter mais informações sobre as CloudWatch métricas que você pode especificar, consulte PredictiveScalingMetricSpecificationa Referência da API do Amazon EC2 Auto Scaling.

Exemplo 1: Uma política de escalabilidade preditiva que cria previsões, mas não implementa a escalabilidade

O exemplo a seguir mostra uma configuração de política completa que usa métricas de utilização da CPU para escalabilidade preditiva com uma utilização-alvo de `40`. O modo `ForecastOnly` é usado por padrão, a menos que você especifique explicitamente qual modo usar. Salve esta configuração em um arquivo chamado `config.json`.

```
{
    "MetricSpecifications": [
        {
            "TargetValue": 40,
            "PredefinedMetricPairSpecification": {
                "PredefinedMetricType": "ASGCPUUtilization"
            }
```

```
        }
    ]
}
```

Para criar a política na linha de comando, execute o put-scaling-policycomando com o arquivo de configuração especificado, conforme demonstrado no exemplo a seguir.

```
aws autoscaling put-scaling-policy --policy-name cpu40-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

Se bem-sucedido, esse comando gerará o nome do recurso da Amazon (ARN) da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/cpu40-predictive-scaling-
policy",
  "Alarms": []
}
```

Exemplo 2: Uma política de escalabilidade preditiva que cria previsões e implementa a escalabilidade

Para uma política que permite que o Amazon EC2 Auto Scaling preveja e implemente a escalabilidade, adicione a propriedade `Mode` com um valor `ForecastAndScale`. O exemplo a seguir mostra uma configuração de política que usa métricas de número de solicitações do Application Load Balancer. A utilização-alvo é `1000` e a escalabilidade preditiva é definida no modo `ForecastAndScale`.

```
{
    "MetricSpecifications": [
        {
            "TargetValue": 1000,
            "PredefinedMetricPairSpecification": {
                "PredefinedMetricType": "ALBRequestCount",
                "ResourceLabel": "app/my-alb/778d41231b141a0f/targetgroup/my-alb-
target-group/943f017f100becff"
            }
        }
    ],
    "Mode": "ForecastAndScale"
}
```

Para criar essa política, execute o put-scaling-policycomando com o arquivo de configuração especificado, conforme demonstrado no exemplo a seguir.

```
aws autoscaling put-scaling-policy --policy-name alb1000-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

Se bem-sucedido, esse comando gerará o nome do recurso da Amazon (ARN) da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-
id:scalingPolicy:19556d63-7914-4997-8c81-d27ca5241386:autoScalingGroupName/my-
asg:policyName/alb1000-predictive-scaling-policy",
  "Alarms": []
}
```

Exemplo 3: Uma política de escalabilidade preditiva que pode escalar acima da capacidade máxima

O exemplo a seguir mostra como criar uma política que poderá escalar além do limite máximo de tamanho do grupo quando você precisar que ele lide com uma carga maior do que o normal. Por padrão, o Amazon EC2 Auto Scaling não aumenta a capacidade do EC2 além da capacidade máxima definida. No entanto, pode ser útil deixá-lo ir além com um pouco mais de capacidade para evitar problemas de performance ou disponibilidade.

Para fornecer espaço para o Amazon EC2 Auto Scaling provisionar capacidade adicional quando a capacidade estiver no tamanho máximo do grupo ou muito próxima a ele, especifique as propriedades `MaxCapacityBreachBehavior` e `MaxCapacityBuffer`, conforme mostrado no exemplo a seguir. É necessário especificar `MaxCapacityBreachBehavior` com um valor de `IncreaseMaxCapacity`. O número máximo de instâncias que seu grupo pode ter depende do valor de `MaxCapacityBuffer`.

```
{
    "MetricSpecifications": [
        {
            "TargetValue": 70,
            "PredefinedMetricPairSpecification": {
                "PredefinedMetricType": "ASGCPUUtilization"
            }
        }
    ],
```

```
    "MaxCapacityBreachBehavior": "IncreaseMaxCapacity",
    "MaxCapacityBuffer": 10
}
```

Neste exemplo, a política é configurada para usar um buffer de 10% (`"MaxCapacityBuffer": 10`). Assim, se a capacidade prevista for 50 e a capacidade máxima for 40, a capacidade máxima efetiva será 55. Uma política que pudesse escalar a capacidade acima da capacidade máxima para igualar, mas não exceder, a capacidade prevista teria um buffer de 0 (`"MaxCapacityBuffer": 0`).

Para criar essa política, execute o put-scaling-policycomando com o arquivo de configuração especificado, conforme demonstrado no exemplo a seguir.

```
aws autoscaling put-scaling-policy --policy-name cpu70-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

Se bem-sucedido, esse comando gerará o nome do recurso da Amazon (ARN) da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:d02ef525-8651-4314-
bf14-888331ebd04f:autoScalingGroupName/my-asg:policyName/cpu70-predictive-scaling-
policy",
  "Alarms": []
}
```

Avaliar as políticas de escalabilidade preditiva

Antes de usar uma política de escalabilidade preditiva para escalar seu grupo do Auto Scaling, analise as recomendações e outros dados da política no console do Amazon EC2 Auto Scaling. Isso é importante porque você não quer que uma política de escalabilidade preditiva escale sua capacidade real até saber que suas previsões estão corretas.

Se o grupo do Auto Scaling for novo, permita que o Amazon EC2 Auto Scaling tenha 24 horas para criar a primeira previsão.

Ao criar uma previsão, o Amazon EC2 Auto Scaling usa dados históricos. Se o grupo do Auto Scaling ainda não tiver muitos dados históricos recentes, o Amazon EC2 Auto Scaling poderá preencher temporariamente a previsão com agregados criados com base nos agregados históricos disponíveis atualmente. As previsões são preenchidas até duas semanas anteriores à data de criação da política.

Conteúdo

- Visualizar recomendações de escalabilidade preditiva
- Analisar grafos de monitoramento de escalabilidade preditiva
- Monitore métricas de escalabilidade preditiva com CloudWatch

Visualizar recomendações de escalabilidade preditiva

Para realizar uma análise eficaz, o Amazon EC2 Auto Scaling deve ter pelo menos duas políticas de escalabilidade preditiva para comparação. (Porém, ainda é possível analisar as conclusões de uma única política.) Ao criar várias políticas, é possível avaliar uma política que usa uma métrica em relação a uma política que usa outra métrica. Também é possível avaliar o impacto de diferentes combinações de valores de destino e métricas. Depois que as políticas de escalabilidade preditiva são criadas, o Amazon EC2 Auto Scaling imediatamente começa a avaliar qual política faria um trabalho melhor ao escalar seu grupo.

Para visualizar as recomendações no console do Amazon EC2 Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Auto Scaling, em Políticas de escalabilidade preditiva, visualize detalhes sobre uma política junto com nossa recomendação. A recomendação indica se é melhor usar a política de escalabilidade preditiva ou não usá-la.

   Se você não tiver certeza se uma política de escalabilidade preditiva é apropriada para seu grupo, analise as colunas Impacto na disponibilidade e Impacto no custo para escolher a política certa. As informações de cada coluna indicam qual é o impacto da política.

   - Impacto na disponibilidade: descreve se a política evitaria um impacto negativo na disponibilidade ao provisionar instâncias suficientes para lidar com a workload, em comparação com o não uso da política.
   - Impacto no custo: descreve se a política evitaria um impacto negativo em seus custos ao não superprovisionar as instâncias, em comparação com o não uso da política. Com o provisionamento excessivo, suas instâncias ficam subutilizadas ou ociosas, o que só aumenta o impacto nos custos.

Se você tiver várias políticas, uma etiqueta Melhor previsão será exibida ao lado do nome da política que oferece mais benefícios de disponibilidade a um custo menor. O impacto na disponibilidade tem um peso maior.

4. (Opcional) Para selecionar o período desejado para os resultados da recomendação, escolha o valor de sua preferência no menu suspenso Período de avaliação: 2 dias, 1 semana, 2 semanas, 4 semanas, 6 semanas ou 8 semanas. Por padrão, o período de avaliação são as duas últimas semanas. Um período de avaliação mais longo oferece mais pontos de dados para os resultados da recomendação. Porém, adicionar mais pontos de dados pode não melhorar os resultados, se seus padrões de carga tiverem sido alterados, como após um período de demanda excepcional. Nesse caso, é possível obter uma recomendação mais focada analisando dados mais recentes.

> **Note**
>
> As recomendações são geradas somente para políticas que estão no modo Somente previsão. O recurso de recomendações funciona melhor quando uma política está no modo Somente previsão durante o período de avaliação. Se você iniciar uma política no modo de Prever e escalar e alterná-la para o modo Somente previsão posteriormente, é provável que as conclusões dessa política tenham desvios. Isso ocorre porque a política já contribuiu em favor da capacidade real.

Analisar grafos de monitoramento de escalabilidade preditiva

No console do Amazon EC2 Auto Scaling, é possível analisar a previsão dos dias, semanas ou meses anteriores para visualizar a performance da política ao longo do tempo. Você também pode usar essas informações para avaliar a precisão das previsões ao decidir se permitirá que uma política escale a capacidade real.

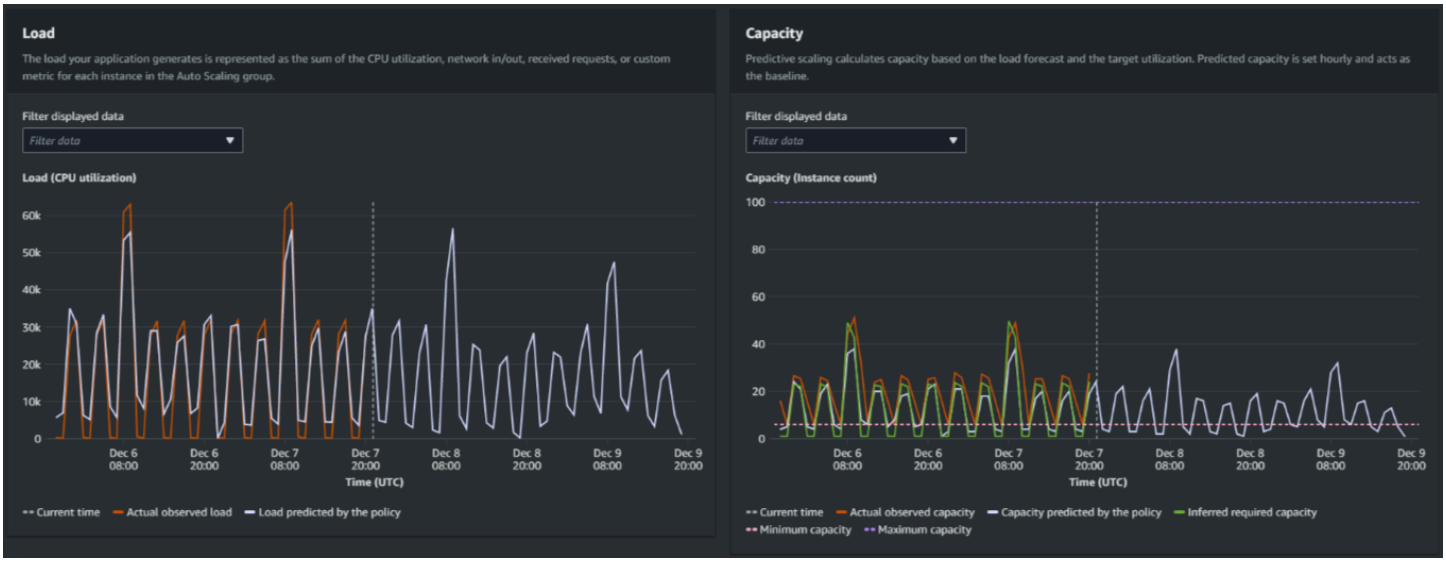
Para analisar grafos de monitoramento de escalabilidade preditiva no console do Amazon EC2 Auto Scaling

1. Escolha uma política na lista Políticas de escalabilidade preditiva.
2. Na seção Monitorar, você pode visualizar as previsões passadas e futuras de carga e de capacidade da política em relação aos valores reais. O grafo Carga exibe a previsão de carga e os valores reais para a métrica de carga escolhida. O grafo Capacidade exibe o número de

instâncias previstas pela política. Também inclui o número real de instâncias iniciadas. A linha vertical separa os valores históricos das previsões futuras. Esses grafos ficam disponíveis logo após a criação da política.

3. (Opcional) Para alterar a quantidade de dados históricos exibidos no gráfico, escolha o valor de sua preferência no menu suspenso Período de avaliação na parte superior da página. O período de avaliação não transforma os dados desta página de maneira alguma. Ele altera apenas a quantidade de dados históricos exibidos.

A imagem a seguir exibe os grafos Carga e Capacidade quando as previsões foram aplicadas várias vezes. A escalabilidade preditiva prevê a carga com base nos dados históricos de carga. A carga que sua aplicação gera é representada como a soma da utilização da CPU, entrada/saída da rede, solicitações recebidas ou métrica personalizada para cada instância no grupo do Auto Scaling. A escalabilidade preditiva calcula as necessidades futuras de capacidade com base na previsão de carga e na utilização desejada que você deseja alcançar para a métrica de escalabilidade.



Compare dados no grafo Carga

Cada linha horizontal representa um conjunto diferente de pontos de dados relatados em intervalos de uma hora:

1. A Carga real observada usa a estatística SUM da métrica de carga escolhida para exibir a carga horária total passada.
2. A Carga prevista pela política exibe a previsão de carga horária. Essa previsão é baseada nas duas semanas anteriores de observações da carga real.

Compare dados no grafo Capacidade

Cada linha horizontal representa um conjunto diferente de pontos de dados relatados em intervalos de uma hora:

1. A capacidade real observada mostra a capacidade real anterior do seu grupo de Auto Scaling quando a GroupTotalInstancesmétrica está ativada. Essa capacidade depende de suas outras políticas de escalabilidade e do tamanho mínimo do grupo durante o período selecionado.
2. A Capacidade prevista pela política exibe a capacidade básica que você pode esperar ter no início de cada hora quando a política estiver no modo de Prever e escalar.
3. A Capacidade necessária inferida exibe a capacidade ideal para manter a métrica de escalabilidade no valor de destino que você escolheu.
4. A Capacidade mínima exibe a capacidade mínima do grupo do Auto Scaling.
5. A Capacidade máxima exibe a capacidade máxima do grupo do Auto Scaling.

Com o objetivo de calcular a capacidade necessária inferida, começamos supondo que cada instância é igualmente utilizada em um valor de destino especificado. Na prática, as instâncias não são utilizadas igualmente. Porém, ao supor que a utilização é distribuída uniformemente entre as instâncias, podemos fazer uma estimativa da probabilidade da quantidade de capacidade necessária. Então, o requisito de capacidade é calculado de modo a ser inversamente proporcional à métrica de escalabilidade que você usou para sua política de escalabilidade preditiva. Em outras palavras, à medida que a capacidade aumenta, a métrica de escalabilidade diminui na mesma proporção. Por exemplo, se a capacidade dobra, a métrica de escalabilidade deve diminuir pela metade.

A fórmula da capacidade necessária inferida:

```
sum of (actualCapacityUnits*scalingMetricValue)/(targetUtilization)
```

Por exemplo, pegamos `actualCapacityUnits` (`10`) e `scalingMetricValue` (`30`) para determinada hora. Em seguida, pegamos a `targetUtilization` especificada em sua política de escalabilidade preditiva (`60`) e calculamos a capacidade necessária inferida para a mesma hora. Isso retorna um valor de `5`. Isso significa que cinco é a quantidade inferida de capacidade necessária para manter a capacidade em proporção inversa direta ao valor de destino da métrica de escala.

> **Note**
>
> Há várias alavancas disponíveis para você ajustar e melhorar a economia de custos e a disponibilidade de sua aplicação.
>
> - Utilize escalabilidade preditiva para a capacidade de linha de base e escalabilidade dinâmica para lidar com capacidade adicional. A escalabilidade dinâmica funciona independentemente da escalabilidade preditiva, aumentando e reduzindo a escala horizontalmente com base na utilização atual. Primeiro, o Amazon EC2 Auto Scaling calcula o número recomendado de instâncias para cada política de escalabilidade dinâmica. Em seguida, ele escala com base na política que fornece o maior número de instâncias.
> - Para permitir que a redução da escala horizontalmente ocorra quando a carga diminui, seu grupo do Auto Scaling deve sempre ter pelo menos uma política de escalabilidade dinâmica com a parte de redução da escala horizontalmente habilitada.
> - Você pode melhorar a performance da escalabilidade verificando se a capacidade mínima e máxima não são muito restritivas. Uma política com um número recomendado de instâncias que não esteja dentro da faixa de capacidade mínima e máxima será impedida de aumentar e reduzir a escala horizontalmente.

Monitore métricas de escalabilidade preditiva com CloudWatch

Dependendo de suas necessidades, talvez você prefira acessar dados de monitoramento para escalabilidade preditiva da Amazon CloudWatch em vez do console Amazon EC2 Auto Scaling. Após criar uma política de escalabilidade preditiva, a política coletará dados que são usados para prever sua carga e capacidade futuras. Depois que esses dados são coletados, eles são armazenados automaticamente em CloudWatch intervalos regulares. Em seguida, você pode usar CloudWatch para visualizar o desempenho da política ao longo do tempo. Você também pode criar CloudWatch alarmes para notificá-lo quando os indicadores de desempenho mudarem além dos limites definidos em CloudWatch.

Tópicos

- Visualizar dados históricos de previsão
- Criar métricas de precisão usando matemática métrica

Visualizar dados históricos de previsão

Você pode visualizar os dados de previsão de carga e capacidade para uma política de escalabilidade preditiva em. CloudWatch Isso pode ser útil ao visualizar previsões em relação a outras CloudWatch métricas em um único gráfico. O recurso também pode ajudar ao visualizar um intervalo mais amplo de tempo para que você possa ver tendências ao longo do tempo. É possível acessar até 15 meses de métricas históricas a fim de obter uma melhor visão do desempenho da política.

Para obter mais informações, consulte Métricas e dimensões de escalabilidade preditiva.

Para visualizar dados históricos de previsão usando o CloudWatch console

1. Abra o CloudWatch console em https://console.aws.amazon.com/cloudwatch/.
2. No painel de navegação, escolha Metrics (Métricas) e, em seguida, All metrics (Todas as métricas).
3. Selecione o namespace da métrica Auto Scaling (Escalabilidade automática).
4. Escolha uma das seguintes opções para visualizar a previsão de carga ou as métricas de previsão de capacidade:
   - Previsões de carga de escalabilidade preditiva
   - Previsões de capacidade de escalabilidade preditiva
5. No campo de pesquisa, insira o nome da política de escalabilidade preditiva ou o nome do grupo do Auto Scaling e pressione a tecla Enter para filtrar os resultados.
6. Para criar um gráfico de uma métrica, marque a caixa de seleção ao lado da métrica. Para alterar o nome do gráfico, escolha o ícone de lápis. Para alterar o período, selecione um dos valores predefinidos ou escolha custom (personalizado). Para obter mais informações, consulte Representação gráfica de uma métrica no Guia do CloudWatch usuário da Amazon.
7. Para alterar a estatística, escolha a guia Métricas em gráfico. Escolha o cabeçalho de coluna ou um valor individual e, em seguida, escolha uma estatística diferente. Embora você possa escolher qualquer estatística para cada métrica, nem todas as estatísticas são úteis para PredictiveScalingLoadForecastas PredictiveScalingCapacityForecastmétricas. Por exemplo, as estatísticas Média, Mínimo e Máximo são úteis, mas a estatística Soma não.
8. Para adicionar outra métrica ao gráfico, em Browse (Procurar), escolha All (Todas), encontre a métrica específica e marque a caixa de seleção ao lado dela. Adicione até 10 métricas.

Por exemplo, para adicionar os valores efetivos de utilização de CPU ao gráfico, escolha o namespace EC2 e, em seguida, escolha By Auto Scaling Group (Por grupo do Auto Scaling). Em seguida, marque a caixa de seleção da CPUUtilizationmétrica e do grupo específico do Auto Scaling.

9. (Opcional) Para adicionar o gráfico a um CloudWatch painel, escolha Ações, Adicionar ao painel.

Criar métricas de precisão usando matemática métrica

Com a matemática métrica, você pode consultar várias CloudWatch métricas e usar expressões matemáticas para criar novas séries temporais com base nessas métricas. Você pode visualizar as séries temporais resultantes no CloudWatch console e adicioná-las aos painéis. Para obter mais informações sobre matemática métrica, consulte Usando matemática métrica no Guia CloudWatch do usuário da Amazon.

Usando matemática métrica, você pode representar graficamente de diferentes maneiras os dados que o Amazon EC2 Auto Scaling gera sobre escalabilidade preditiva. Isso ajuda a monitorar o desempenho da política ao longo do tempo e a entender se é possível melhorar sua combinação de métricas.

Por exemplo, você pode usar uma expressão de matemática métrica para monitorar o mean absolute percentage error (MAPE – Erro percentual absoluto médio). A métrica MAPE ajuda a monitorar a diferença entre os valores previstos e os valores efetivos observados durante uma determinada janela de previsão. Mudanças no valor de MAPE podem indicar se o desempenho da política está se degradando ao longo do tempo conforme a natureza da sua aplicação muda. Um aumento no MAPE sinaliza uma lacuna maior entre os valores previstos e os valores efetivos.

Exemplo: expressão de matemática métrica

Para começar a usar esse tipo de gráfico, você pode criar uma expressão de matemática métrica como a mostrada no exemplo a seguir.

```
{
  "MetricDataQueries": [
    {
      "Expression": "TIME_SERIES(AVG(ABS(m1-m2)/m1))",
      "Id": "e1",
      "Period": 3600,
      "Label": "MeanAbsolutePercentageError",
      "ReturnData": true
```

```
    },
    {
      "Id": "m1",
      "Label": "ActualLoadValues",
      "MetricStat": {
        "Metric": {
          "Namespace": "AWS/EC2",
          "MetricName": "CPUUtilization",
          "Dimensions": [
            {
              "Name": "AutoScalingGroupName",
              "Value": "my-asg"
            }
          ]
        },
        "Period": 3600,
        "Stat": "Sum"
      },
      "ReturnData": false
    },
    {
      "Id": "m2",
      "Label": "ForecastedLoadValues",
      "MetricStat": {
        "Metric": {
          "Namespace": "AWS/AutoScaling",
          "MetricName": "PredictiveScalingLoadForecast",
          "Dimensions": [
            {
              "Name": "AutoScalingGroupName",
              "Value": "my-asg"
            },
            {
              "Name": "PolicyName",
              "Value": "my-predictive-scaling-policy"
            },
            {
              "Name": "PairIndex",
              "Value": "0"
            }
          ]
        },
        "Period": 3600,
        "Stat": "Average"
```

```
      },
      "ReturnData": false
    }
  ]
}
```

Em vez de uma só métrica, há uma matriz de estruturas de consulta de dados métricos para `MetricDataQueries`. Cada item em `MetricDataQueries` obtém uma métrica ou executa uma expressão matemática. O primeiro item, `e1`, é a expressão matemática. A expressão designada define o parâmetro `ReturnData` como `true`, resultando na produção de uma única série temporal. Para todas as outras métricas, o valor `ReturnData` é `false`.

No exemplo, a expressão designada usa os valores reais e previstos como entrada e retorna a nova métrica (MAPE). `m1`é a CloudWatch métrica que contém os valores reais de carga (supondo que a utilização da CPU seja a métrica de carga originalmente especificada para a política nomeada`my-predictive-scaling-policy`). `m2`é a CloudWatch métrica que contém os valores de carga previstos. A sintaxe matemática para a métrica MAPE é a seguinte:

média de (abs ((efetivo - previsto)/(efetivo)))

Visualizar suas métricas de precisão e definir alarmes

Para visualizar os dados métricos de precisão, selecione a guia Métricas no CloudWatch console. Nele, é possível representar graficamente os dados. Para obter mais informações, consulte Adicionar uma expressão matemática a um CloudWatch gráfico no Guia do CloudWatch usuário da Amazon.

Na seção Metrics (Métricas), você também pode criar um alarme com base em uma métrica que esteja monitorando. Enquanto estiver na guia Graphed metrics (Métricas representadas em gráficos), selecione o ícone Create alarm (Criar alarme) na coluna Actions (Ações). O ícone Create alarm (Criar alarme) é representado como um pequeno sino. Para obter mais informações e opções de notificação, consulte Criação de um CloudWatch alarme com base em uma expressão matemática métrica e Notificação de usuários sobre alterações de alarme no Guia do CloudWatch usuário da Amazon.

Como alternativa, você pode usar GetMetricDatae PutMetricAlarmrealizar cálculos usando matemática métrica e criar alarmes com base na saída.

Substituir valores de previsão usando ações programadas

Às vezes, você pode ter informações adicionais sobre seus futuros requisitos de aplicações que o cálculo de previsão não pode levar em conta. Por exemplo, os cálculos de previsão podem

subestimar a capacidade necessária para um evento de marketing futuro. Você pode usar ações programadas para substituir temporariamente a previsão durante períodos futuros. As ações programadas podem ser executadas de forma recorrente ou em uma data e hora específicas quando houver flutuações de demanda únicas.

Por exemplo, você pode criar uma ação programada com uma capacidade mínima maior do que a prevista. Em runtime, o Amazon EC2 Auto Scaling atualiza a capacidade mínima do grupo do Auto Scaling. Como a escalabilidade preditiva otimiza a capacidade, uma ação agendada com uma capacidade mínima maior que os valores de previsão é honrada. Isso impede que a capacidade seja menor do que o esperado. Para interromper a substituição da previsão, use uma segunda ação programada para retornar a capacidade mínima à configuração original.

O procedimento a seguir descreve as etapas necessárias para substituir a previsão durante períodos futuros.

Tópicos

- Etapa 1: (Opcional) Analisar dados de séries temporais
- Etapa 2: Criar duas ações programadas

> **Important**
>
> Este tópico pressupõe que você esteja tentando substituir a previsão para escalar para uma capacidade maior do que a prevista. Se você precisar diminuir temporariamente a capacidade sem interferência de uma política de escala preditiva, use o modo somente de previsão. Enquanto estiver no modo somente de previsão, a escala preditiva continuará a gerar previsões, mas não aumentará automaticamente a capacidade. Em seguida, você pode monitorar a utilização dos recursos e diminuir manualmente o tamanho do grupo, conforme necessário. Para obter mais informações sobre a escalabilidade manual, consulte Escalabilidade manual para o Amazon EC2 Auto Scaling.

Etapa 1: (Opcional) Analisar dados de séries temporais

Comece analisando os dados de séries temporais de previsão. Essa é uma etapa opcional, mas é útil quando você deseja entender os detalhes da previsão.

1. Recuperar a previsão

Após a criação da previsão, é possível consultar um período específico na previsão. O objetivo da consulta é obter uma visão completa dos dados de séries temporais para um período específico.

Sua consulta pode incluir até dois dias de dados de previsão futura. Se você usa a escalabilidade preditiva há algum tempo, também pode acessar seus dados de previsão anteriores. No entanto, a duração máxima de tempo entre as horas inicial e final é de 30 dias.

Para obter a previsão usando o get-predictive-scaling-forecast AWS CLI comando, forneça os seguintes parâmetros no comando:

- Insira o nome do grupo do Auto Scaling no parâmetro `--auto-scaling-group-name`.
- Insira o nome da política no parâmetro `--policy-name`.
- Insira a hora de início no parâmetro `--start-time` para retornar apenas os dados de previsão para depois ou no horário especificado.
- Insira a hora de término no parâmetro `--end-time` para retornar apenas os dados de previsão para antes do horário especificado.

```
aws autoscaling get-predictive-scaling-forecast --auto-scaling-group-name my-asg \
    --policy-name cpu40-predictive-scaling-policy \
    --start-time "2021-05-19T17:00:00Z" \
    --end-time "2021-05-19T23:00:00Z"
```

Se bem-sucedido, o comando retornará uma resposta semelhante à seguinte.

```
{
    "LoadForecast": [
        {
            "Timestamps": [
                "2021-05-19T17:00:00+00:00",
                "2021-05-19T18:00:00+00:00",
                "2021-05-19T19:00:00+00:00",
                "2021-05-19T20:00:00+00:00",
                "2021-05-19T21:00:00+00:00",
                "2021-05-19T22:00:00+00:00",
                "2021-05-19T23:00:00+00:00"
            ],
            "Values": [
```

```
                153.0655799339254,
                128.8288551285919,
                107.1179447150675,
                197.3601844551528,
                626.4039934516954,
                596.9441277518481,
                677.9675713779869
            ],
            "MetricSpecification": {
                "TargetValue": 40.0,
                "PredefinedMetricPairSpecification": {
                    "PredefinedMetricType": "ASGCPUUtilization"
                }
            }
        }
    ],
    "CapacityForecast": {
        "Timestamps": [
            "2021-05-19T17:00:00+00:00",
            "2021-05-19T18:00:00+00:00",
            "2021-05-19T19:00:00+00:00",
            "2021-05-19T20:00:00+00:00",
            "2021-05-19T21:00:00+00:00",
            "2021-05-19T22:00:00+00:00",
            "2021-05-19T23:00:00+00:00"
        ],
        "Values": [
            2.0,
            2.0,
            2.0,
            2.0,
            4.0,
            4.0,
            4.0
        ]
    },
    "UpdateTime": "2021-05-19T01:52:50.118000+00:00"
}
```

A resposta inclui duas previsões: `LoadForecast` e `CapacityForecast`. `LoadForecast` mostra a previsão de carga horária. `CapacityForecast` mostra os valores de previsão para a capacidade que é necessária em uma base horária para lidar com a carga prevista enquanto mantém um `TargetValue` de 40,0 (40% de utilização média da CPU).

2. Identificar o período-alvo

   Identifique a hora ou horas em que a flutuação de demanda única deverá ocorrer. Lembre-se de que as datas e os horários mostrados na previsão estão em UTC.

Etapa 2: Criar duas ações programadas

Em seguida, crie duas ações programadas para um período específico em que sua aplicação terá uma carga maior do que a prevista. Por exemplo, se você tiver um evento de marketing que irá direcionar o tráfego para seu site por um período limitado, poderá programar uma ação única para atualizar a capacidade mínima quando ele começar. Em seguida, agende outra ação para retornar a capacidade mínima para a configuração original quando o evento terminar.

Para criar duas ações programadas para eventos únicos (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Automatic scaling (Escalabilidade automática), em Scheduled actions (Ações programadas), escolha Create scheduled action (Criar ação programada).
4. Preencha as seguintes configurações de ações programadas:

   a. Insira um Name (Nome) para a ação programada.

   b. Em Min (Mínima) insira a nova capacidade mínima para seu grupo do Auto Scaling. A capacidade Min (Mínima) deve ser menor ou igual ao tamanho máximo do grupo. Se o valor de Min (Mínima) for maior que o tamanho máximo do grupo, será necessário atualizar o valor de Max (Máxima).

   c. Em Recurrence (Recorrência), escolha Once (Uma vez).

   d. Em Time zone (Fuso horário), escolha um fuso horário. Se nenhum fuso horário for escolhido, `ETC/UTC` será usado por padrão.

   e. Defina uma Specific start time (Hora de início específica).
5. Escolha Criar.

   O console exibe as ações programadas para o grupo do Auto Scaling.

6. Configure uma segunda ação programada para retornar a capacidade mínima para a configuração original no final do evento. A escalabilidade preditiva pode escalar a capacidade somente quando o valor definido para Min (Mínima) é menor que os valores da previsão.

Para criar duas ações programadas para eventos únicos (AWS CLI)

Para usar o AWS CLI para criar as ações agendadas, use o comando put-scheduled-update-group-action.

Por exemplo, vamos definir uma programação que mantenha uma capacidade mínima de três instâncias em 19 de maio às 17h por oito horas. Os comandos a seguir mostram como implementar esse cenário.

O primeiro comando put-scheduled-update-group-action instrui o Amazon EC2 Auto Scaling a atualizar a capacidade mínima do grupo de Auto Scaling especificado às 17h UTC de 19 de maio de 2021.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-event-start \
  --auto-scaling-group-name my-asg --start-time "2021-05-19T17:00:00Z" --minimum-capacity 3
```

O segundo comando instrui o Amazon EC2 Auto Scaling a definir a capacidade mínima do grupo como um à 1h da manhã UTC em 20 de maio de 2021.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-event-end \
  --auto-scaling-group-name my-asg --start-time "2021-05-20T01:00:00Z" --minimum-capacity 1
```

Após você adicionar essas ações programadas ao grupo do Auto Scaling, o Amazon EC2 Auto Scaling fará o seguinte:

- Às 17h UTC em 19 de maio de 2021, a primeira ação programada é executada. Se o grupo tiver menos de três instâncias, ele será expandido para três instâncias. Durante esse período e nas próximas oito horas, o Amazon EC2 Auto Scaling poderá continuar a aumentar a escala na horizontal se a capacidade prevista for maior do que a capacidade real ou se houver uma política de escalabilidade dinâmica em vigor.

- À 1h da manhã UTC em 20 de maio de 2021, a segunda ação programada é executada. Isso retorna a capacidade mínima para sua configuração original no final do evento.

Escalabilidade com base em programações recorrentes

Para substituir a previsão para o mesmo período de tempo todas as semanas, crie duas ações programadas e forneça a lógica de hora e data usando uma expressão cron.

A expressão cron consiste em cinco campos separados por espaços: [Minute] [Hour] [Day_of_Month] [Month_of_Year] [Day_of_Week]. Os campos podem conter quaisquer valores permitidos, incluindo caracteres especiais.

Por exemplo, esta expressão cron executa a ação todas as terças-feiras às 6h30. O asterisco é usado como um curinga para corresponder a todos os valores de um campo.

```
30 6 * * 2
```

Consulte também

Para obter mais informações sobre como criar, listar, editar e excluir ações programadas, consulte Escalabilidade programada para o Amazon EC2 Auto Scaling.

Configurações avançadas de política de escalabilidade preditiva usando métricas personalizadas

Em uma política de escalabilidade preditiva é possível usar métricas predefinidas ou personalizadas. Métricas personalizadas são úteis quando as métricas predefinidas (CPU, E/S da rede e contagem de solicitações do Application Load Balancer) não descrevem suficientemente a carga da aplicação.

Ao criar uma política de escalabilidade preditiva com métricas personalizadas, você pode especificar outras CloudWatch métricas fornecidas por AWS, ou você pode especificar métricas que você mesmo define e publica. Você também pode usar a matemática métrica para agregar e transformar métricas existentes em uma nova série temporal que AWS não é rastreada automaticamente.
A combinação de valores em seus dados, por exemplo, calculando novas somas ou médias, é chamada de agregação. Os dados resultantes são chamados de um agregado.

A seção a seguir contém as práticas recomendados e exemplos de como sstruturar o JSON para a política.

Tópicos

- Práticas recomendadas
- Pré-requisitos
- Estruture o JSON para métricas personalizadas
- Criar uma política de escalação preditiva no console (métricas personalizadas)
- Limitações

Práticas recomendadas

As seguintes práticas recomendadas podem ajudar no uso mais eficaz de métricas personalizadas:

- Para a especificação da métrica de carga, a métrica mais útil é uma métrica que represente a carga em um grupo do Auto Scaling como um todo, independentemente da capacidade do grupo.
- Para a especificação da métrica de escalabilidade, a métrica mais útil para escalar é throughput ou utilização média por métrica de instância.
- A métrica de escalabilidade deve ser inversamente proporcional à capacidade. Ou seja, se o número de instâncias no grupo do Auto Scaling aumentar, a métrica de escalabilidade deve diminuir aproximadamente na mesma proporção. Para garantir que a escalabilidade preditiva se comporte conforme o esperado, a métrica de carga e a métrica de escalabilidade também devem se correlacionar fortemente entre si.
- A utilização visada deve corresponder ao tipo de métrica de escalabilidade. Para uma configuração de política que use a utilização da CPU, essa é uma porcentagem visada. Para uma configuração de política que use throughput, como o número de solicitações ou mensagens, esse é o número visado de solicitações ou mensagens por instância durante qualquer intervalo de um minuto.
- Se essas recomendações não forem seguidas, provavelmente os valores futuros previstos da série temporal estarão incorretos. Para validar se os dados estão corretos, você pode visualizar os valores previstos no console do Amazon EC2 Auto Scaling. Como alternativa, depois de criar sua política de escalabilidade preditiva, inspecione os `CapacityForecast` objetos `LoadForecast` e os objetos retornados por uma chamada para a API. GetPredictiveScalingForecast
- Recomendamos a configuração da escalabilidade preditiva no modo apenas previsão para avaliar a previsão antes que a escalabilidade preditiva comece a modificar ativamente a capacidade.

Pré-requisitos

Para adicionar métricas personalizadas à política de escalação preditiva, é necessário ter as permissões `cloudwatch:GetMetricData`.

Para especificar suas próprias métricas em vez das métricas AWS fornecidas, você deve primeiro publicar suas métricas em CloudWatch. Para obter mais informações, consulte Publicação de métricas personalizadas no Guia CloudWatch do usuário da Amazon.

Se publicar suas próprias métricas, certifique-se de publicar os pontos de dados com uma frequência mínima de cinco minutos. O Amazon EC2 Auto Scaling recupera os CloudWatch pontos de dados com base na duração do período necessário. Por exemplo, a especificação da métrica de carga usa métricas horárias para medir a carga em seu aplicativo. CloudWatch usa seus dados métricos publicados para fornecer um único valor de dados para qualquer período de uma hora, agregando todos os pontos de dados com registros de data e hora que se enquadram em cada período de uma hora.

Estruture o JSON para métricas personalizadas

A seção a seguir contém exemplos de como configurar a escala preditiva para consultar dados. CloudWatch Há dois métodos diferentes de configurar essa opção, e o método escolhido afeta qual será o formato usado para estruturar JSON para a política de escalação preditiva. Quando você usa matemática de métricas, o formato do JSON varia ainda mais com base na matemática de métrica que está sendo aplicada.

1. Para criar uma política que obtenha dados diretamente de outras CloudWatch métricas fornecidas AWS ou nas quais você publica CloudWatch, consulteExemplo de política de escalação preditiva com métricas personalizadas de carga e de dimensionamento (AWS CLI).
2. Para criar uma política que possa consultar várias CloudWatch métricas e usar expressões matemáticas para criar novas séries temporais com base nessas métricas, consulteUsar expressões de matemática métrica.

Exemplo de política de escalação preditiva com métricas personalizadas de carga e de dimensionamento (AWS CLI)

Para criar uma política de escalabilidade preditiva com métricas personalizadas de carga e escalabilidade com o. AWS CLI, armazene os argumentos para `--predictive-scaling-configuration` em um arquivo JSON chamado. `config.json`

Você começa a adicionar métricas personalizadas substituindo os valores substituíveis no exemplo a seguir por suas métricas e sua meta de utilização.

```
{
  "MetricSpecifications": [
```

```
    {
      "TargetValue": 50,
      "CustomizedScalingMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "scaling_metric",
            "MetricStat": {
              "Metric": {
                "MetricName": "MyUtilizationMetric",
                "Namespace": "MyNameSpace",
                "Dimensions": [
                  {
                    "Name": "MyOptionalMetricDimensionName",
                    "Value": "MyOptionalMetricDimensionValue"
                  }
                ]
              },
              "Stat": "Average"
            }
          }
        ]
      },
      "CustomizedLoadMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "load_metric",
            "MetricStat": {
              "Metric": {
                "MetricName": "MyLoadMetric",
                "Namespace": "MyNameSpace",
                "Dimensions": [
                  {
                    "Name": "MyOptionalMetricDimensionName",
                    "Value": "MyOptionalMetricDimensionValue"
                  }
                ]
              },
              "Stat": "Sum"
            }
          }
        ]
      }
    }
  ]
```

```
}
```

Para obter mais informações, consulte a MetricDataQueryReferência da API Amazon EC2 Auto Scaling.

> **Note**
>
> Veja a seguir alguns recursos adicionais que podem ajudar você a encontrar nomes de métricas, namespaces, dimensões e estatísticas para CloudWatch métricas:
>
> - Para obter informações sobre as métricas disponíveis para AWS serviços, consulte AWS serviços que publicam CloudWatch métricas no Guia CloudWatch do usuário da Amazon.
> - Para obter o nome exato da métrica, o namespace e as dimensões (se aplicável) de uma CloudWatch métrica com o AWS CLI, consulte list-metrics.

Para criar essa política, execute o put-scaling-policycomando usando o arquivo JSON como entrada, conforme demonstrado no exemplo a seguir.

```
aws autoscaling put-scaling-policy --policy-name my-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

Se bem-sucedido, esse comando gerará o nome do recurso da Amazon (ARN) da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/my-predictive-scaling-policy",
  "Alarms": []
}
```

Usar expressões de matemática métrica

A seção a seguir fornece informações e exemplos de políticas de escalação preditiva que mostram como você pode usar a matemática de métricas em sua política.

Tópicos

- Noções básicas de matemática métrica

- Exemplo de política de escalação preditiva que combina métricas por meio da matemática de métricas (AWS CLI)
- Exemplo de política de escalabilidade preditiva para usar em um cenário de blue/green implantação ()AWS CLI

Noções básicas de matemática métrica

Se tudo o que você quer fazer é agregar dados métricos existentes, a matemática CloudWatch métrica poupa o esforço e o custo de publicar outra métrica no. CloudWatch Você pode usar qualquer métrica que AWS forneça e também pode usar métricas que você define como parte de seus aplicativos. Por exemplo, talvez você queira calcular a lista de pendências da fila do Amazon SQS por instância. Você pode fazer isso usando o número aproximado de mensagens disponíveis para recuperação da fila e dividindo esse número pela capacidade de execução do grupo do Auto Scaling.

Para obter mais informações, consulte Usando matemática métrica no Guia CloudWatch do usuário da Amazon.

Se você optar por usar uma expressão matemática métrica em sua política de escalabilidade preditiva, considere os seguintes pontos:

- As operações matemáticas métricas usam os pontos de dados da combinação exclusiva de nome da métrica, namespace e pares de métricas de chaves-valor da dimensão.
- Você pode usar qualquer operador aritmético (+ - */^), função estatística (como AVG ou SUM) ou outra função compatível. CloudWatch
- Você pode usar as métricas e os resultados de outras expressões matemáticas nas fórmulas da expressão matemática.
- Suas expressões matemáticas métricas podem ser compostas de agregações diferentes. No entanto, uma prática recomendada para o resultado final da agregação é usar `Average` para a métrica de escalabilidade e `Sum` para a métrica de carga.
- Qualquer expressão usada em uma especificação de métrica deve eventualmente retornar uma única série temporal.

Para usar matemática métrica, faça o seguinte:

- Escolha uma ou mais CloudWatch métricas. Em seguida, crie a expressão. Para obter mais informações, consulte Usando matemática métrica no Guia CloudWatch do usuário da Amazon.

- Verifique se a expressão matemática métrica é válida usando o CloudWatch console ou a CloudWatch GetMetricDataAPI.

Exemplo de política de escalação preditiva que combina métricas por meio da matemática de métricas (AWS CLI)

Às vezes, ao invés de especificar a métrica diretamente, talvez seja necessário processar seus dados de alguma forma, primeiramente. Por exemplo, você pode ter uma aplicação que extrai o trabalho de uma fila do Amazon SQS e talvez queira usar o número de itens na fila como critério para escalabilidade preditiva. O número de mensagens na fila não define unicamente o número necessário de instâncias. Portanto, é necessário mais trabalho para criar uma métrica que possa ser usada para calcular a lista de pendências por instância. Para obter mais informações, consulte Ajuste de escala baseado no Amazon SQS.

Veja a seguir um exemplo de política de escalabilidade preditiva para esse cenário. Ele especifica métricas de escalabilidade e carga baseadas na métrica `ApproximateNumberOfMessagesVisible` do Amazon SQS, que é o número de mensagens disponíveis para recuperação da fila. Ele também usa a métrica `GroupInServiceInstances` do Amazon EC2 Auto Scaling e uma expressão matemática para calcular a lista de pendências por instância para a métrica de escalabilidade.

```
aws autoscaling put-scaling-policy --policy-name my-sqs-custom-metrics-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
{
  "MetricSpecifications": [
    {
      "TargetValue": 100,
      "CustomizedScalingMetricSpecification": {
        "MetricDataQueries": [
          {
            "Label": "Get the queue size (the number of messages waiting to be
 processed)",
            "Id": "queue_size",
            "MetricStat": {
              "Metric": {
                "MetricName": "ApproximateNumberOfMessagesVisible",
                "Namespace": "AWS/SQS",
                "Dimensions": [
                  {
```

```
                "Name": "QueueName",
                "Value": "my-queue"
              }
            ]
          },
          "Stat": "Sum"
        },
        "ReturnData": false
      },
      {
        "Label": "Get the group size (the number of running instances)",
        "Id": "running_capacity",
        "MetricStat": {
          "Metric": {
            "MetricName": "GroupInServiceInstances",
            "Namespace": "AWS/AutoScaling",
            "Dimensions": [
              {
                "Name": "AutoScalingGroupName",
                "Value": "my-asg"
              }
            ]
          },
          "Stat": "Sum"
        },
        "ReturnData": false
      },
      {
        "Label": "Calculate the backlog per instance",
        "Id": "scaling_metric",
        "Expression": "queue_size / running_capacity",
        "ReturnData": true
      }
    ]
  },
  "CustomizedLoadMetricSpecification": {
    "MetricDataQueries": [
      {
        "Id": "load_metric",
        "MetricStat": {
          "Metric": {
            "MetricName": "ApproximateNumberOfMessagesVisible",
            "Namespace": "AWS/SQS",
            "Dimensions": [
```

```
                {
                  "Name": "QueueName",
                  "Value": "my-queue"
                }
              ],
            },
            "Stat": "Sum"
          },
          "ReturnData": true
        }
      ]
    }
  }
]
}
```

O exemplo retorna o ARN da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/my-sqs-custom-metrics-policy",
  "Alarms": []
}
```

Exemplo de política de escalabilidade preditiva para usar em um cenário de blue/green implantação ()AWS CLI

Uma expressão de pesquisa fornece uma opção avançada na qual você pode consultar para obter uma métrica de vários grupos do Auto Scaling e realizar expressões matemáticas neles. Isso é especialmente útil para blue/green implantações.

> **Note**
>
> Uma implantação azul/verde é um método de implantação no qual você cria dois grupos do Auto Scaling separados, mas idênticos. Apenas um dos grupos recebe tráfego de produção. O tráfego do usuário é inicialmente direcionado para o grupo do Auto Scaling anterior (“azul”), enquanto um novo grupo (“verde”) é usado para testar e avaliar uma nova versão de uma aplicação ou serviço. O tráfego do usuário é deslocado para o grupo do Auto Scaling verde depois que uma nova implantação é testada e aceita. Em seguida, é possível excluir o grupo azul depois que a implantação for bem-sucedida.

Quando novos grupos de Auto Scaling são criados como parte de uma blue/green implantação, o histórico métrico de cada grupo pode ser incluído automaticamente na política de escalabilidade preditiva sem que você precise alterar suas especificações métricas. Para obter mais informações, consulte Usando políticas de escalabilidade preditiva do EC2 Auto Scaling com implantações azul/verdes no blog de computação. AWS

O exemplo de política a seguir mostra como isso pode ser feito. Neste exemplo, a política usa a métrica `CPUUtilization` emitida pelo Amazon EC2. Ela também usa a métrica `GroupInServiceInstances` do Amazon EC2 Auto Scaling e uma expressão matemática para calcular o valor da métrica de escalabilidade por instância. Ela também especifica uma especificação de métrica de capacidade para obter a métrica `GroupInServiceInstances`.

A expressão de pesquisa encontra o `CPUUtilization` de instâncias em vários grupos do Auto Scaling com base nos critérios de pesquisa especificados. Se, posteriormente, você criar um novo grupo do Auto Scaling que corresponda aos mesmos critérios de pesquisa, o `CPUUtilization` das instâncias no novo grupo do Auto Scaling são incluídas automaticamente.

```
aws autoscaling put-scaling-policy --policy-name my-blue-green-predictive-scaling-
policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
{
  "MetricSpecifications": [
    {
      "TargetValue": 25,
      "CustomizedScalingMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "load_sum",
            "Expression": "SUM(SEARCH('{AWS/EC2,AutoScalingGroupName} MetricName=
\"CPUUtilization\" ASG-myapp', 'Sum', 300))",
            "ReturnData": false
          },
          {
            "Id": "capacity_sum",
            "Expression": "SUM(SEARCH('{AWS/AutoScaling,AutoScalingGroupName}
 MetricName=\"GroupInServiceInstances\" ASG-myapp', 'Average', 300))",
            "ReturnData": false
          },
          {
            "Id": "weighted_average",
            "Expression": "load_sum / capacity_sum",
```

```
            "ReturnData": true
          }
        ]
      },
      "CustomizedLoadMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "load_sum",
            "Expression": "SUM(SEARCH('{AWS/EC2,AutoScalingGroupName} MetricName=
\"CPUUtilization\" ASG-myapp', 'Sum', 3600))"
          }
        ]
      },
      "CustomizedCapacityMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "capacity_sum",
            "Expression": "SUM(SEARCH('{AWS/AutoScaling,AutoScalingGroupName}
 MetricName=\"GroupInServiceInstances\" ASG-myapp', 'Average', 300))"
          }
        ]
      }
    }
  ]
}
```

O exemplo retorna o ARN da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/my-blue-green-predictive-
scaling-policy",
  "Alarms": []
}
```

Criar uma política de escalação preditiva no console (métricas personalizadas)

Se ocorrer um problema ao usar métricas personalizadas, recomendamos fazer o seguinte:

- Se uma mensagem de erro for fornecida, leia a mensagem e resolva o problema que ela relata, se possível.

- Se ocorrer um problema ao tentar usar uma expressão de pesquisa em um cenário de blue/green implantação, primeiro certifique-se de entender como criar uma expressão de pesquisa que procure uma correspondência parcial em vez de uma correspondência exata. Além disso, confira se sua consulta localiza apenas os grupos do Auto Scaling que estão executando a aplicação específica. Para obter mais informações sobre a sintaxe da expressão de pesquisa, consulte a sintaxe CloudWatch da expressão de pesquisa no Guia CloudWatch do usuário da Amazon.
- Se você não validou uma expressão com antecedência, o put-scaling-policycomando a valida ao criar sua política de escalabilidade. No entanto, existe a possibilidade de que esse comando não identifique a causa exata dos erros detectados. Para corrigir os problemas, solucione os erros que você recebe em uma resposta de uma solicitação ao get-metric-datacomando. Você também pode solucionar o problema da expressão no CloudWatch console.
- Quando você visualiza os gráficos de Carga e de Capacidade no console, o gráfico da Capacidade pode não mostrar nenhum dado. Para garantir que os gráficos tenham dados completos, certifique-se de habilitar consistentemente métricas de grupo para seus grupos do Auto Scaling. Para obter mais informações, consulte Ativar métricas do grupo do Auto Scaling (console).
- A especificação da métrica de capacidade só é útil para implantações azul/verde quando você tem aplicações que são executadas em diferentes grupos do Auto Scaling ao longo de suas vidas úteis. Essa métrica personalizada permite que você forneça a capacidade total de vários grupos do Auto Scaling. A escalabilidade preditiva usa isso para mostrar dados históricos nos gráficos de Capacidade no console.
- É necessário especificar `false` para `ReturnData` se `MetricDataQueries` especificar a função SEARCH() (BUSCAR) por conta própria sem uma função matemática como SUM() (SOMA). Isso ocorre porque as expressões de pesquisa podem retornar várias séries temporais, e uma especificação métrica baseado em uma expressão pode retornar apenas uma série temporal.
- Todas as métricas envolvidas em uma expressão de pesquisa devem ter a mesma resolução.

Limitações

- Você pode consultar pontos de dados de até 10 métricas em uma especificação métrica.
- Para os propósitos desse limite, uma expressão conta como uma métrica.

Controlar quais instâncias do Auto Scaling serão terminadas durante uma redução de escala na horizontal

O Amazon EC2 Auto Scaling usa políticas de terminação para decidir a ordem de terminação das instâncias. Você pode usar uma política predefinida ou criar uma política personalizada para atender a seus requisitos específicos. Ao usar uma política personalizada ou uma escala de instância na proteção, você também pode impedir que o grupo do Auto Scaling termine instâncias que ainda não estão prontas para serem terminadas.

Conteúdo

- Trabalhar com políticas de término do Amazon EC2 Auto Scaling
- Trabalhar com políticas de término do Amazon EC2 Auto Scaling
- Criar uma política de término personalizada com o Lambda
- Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância
- Projete suas aplicações para lidar com a terminação de instâncias sem problemas

Trabalhar com políticas de término do Amazon EC2 Auto Scaling

As seções a seguir descrevem os cenários em que o Amazon EC2 Auto Scaling usa políticas de término.

Conteúdo

- Eventos de redução de escala horizontalmente
- Atualização de instância
- Rebalanceamento de zona de disponibilidade

Eventos de redução de escala horizontalmente

Um evento de redução de escala na horizontal ocorre quando há um novo valor para a capacidade desejada de um grupo do Auto Scaling que é menor do que a capacidade atual do grupo.

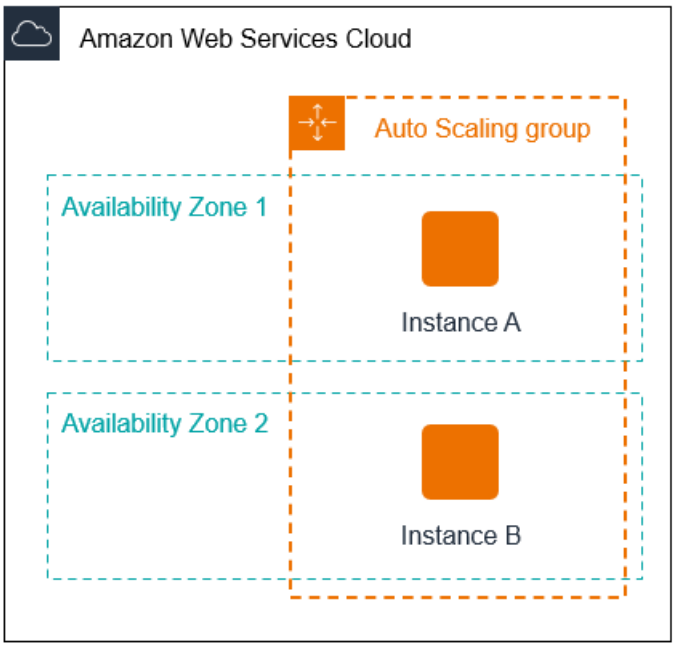
Eventos de redução de escala na horizontal ocorrem nos cenários a seguir:

- Ao usar políticas de escalabilidade dinâmica e o tamanho do grupo diminui como resultado de alterações no valor de uma métrica

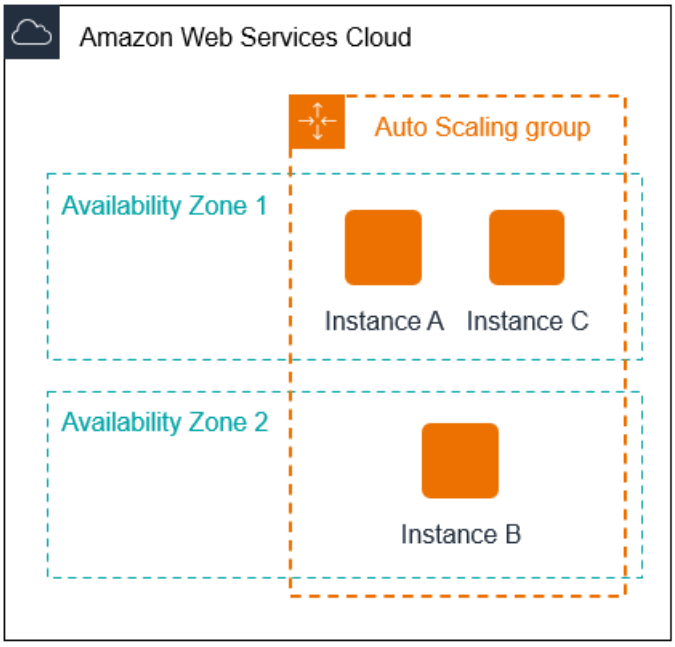
- Ao usar a escalabilidade programada e o tamanho do grupo diminui como resultado de uma ação programada
- Quando você reduz o tamanho do grupo manualmente

O exemplo a seguir mostra como as políticas de término funcionam quando há um evento de redução de capacidade na horizontal.

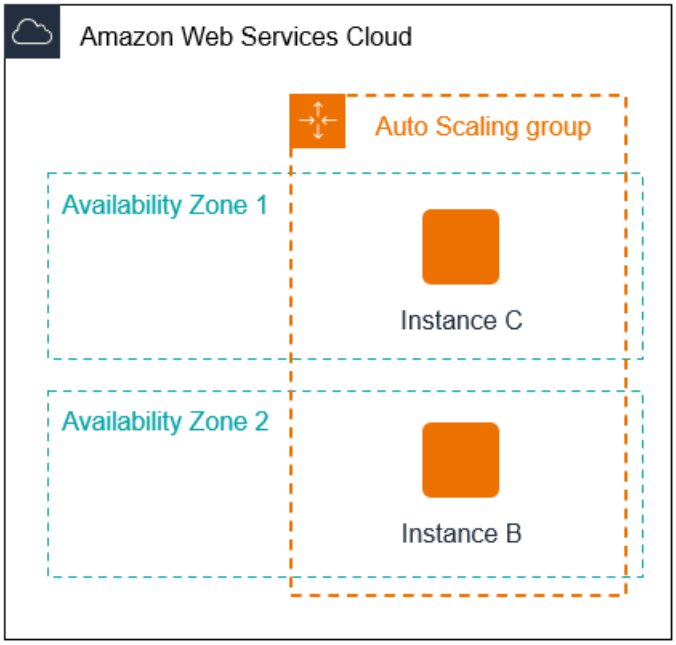
1. O grupo do Auto Scaling deste exemplo tem um tipo de instância, duas zonas de disponibilidade e uma capacidade desejada de duas instâncias. Ele também tem uma política de escalabilidade dinâmica que adiciona e remove instâncias quando a utilização de recursos aumenta ou diminui. As duas instâncias desse grupo são distribuídas nas duas zonas de disponibilidade, como mostrado no diagrama a seguir.



2. Quando o grupo do Auto Scaling aumenta a escala na horizontal, o Amazon EC2 Auto Scaling executa uma nova instância. O grupo do Auto Scaling agora possui três instâncias, distribuídas nas duas zonas de disponibilidade, como mostrado no diagrama a seguir.



3. Quando o grupo do Auto Scaling reduz a escala na horizontal, o Amazon EC2 Auto Scaling termina uma das instâncias.
4. Se você não tiver atribuído uma política de término específica ao grupo, o Amazon EC2 Auto Scaling usará a política de término padrão. Ele selecionará a zona de disponibilidade com duas instâncias e terminará a instância que foi iniciada com o modelo de execução ou a configuração de execução mais antiga. Se as instâncias tiverem sido inidicadas com o mesmo modelo de execução, o Amazon EC2 Auto Scaling selecionará a instância que estiver mais perto da próxima hora de faturamento e a terminará.



Atualização de instância

Você inicia as atualizações de instâncias para atualizar as instâncias em seu grupo do Auto Scaling. Durante uma atualização de instância, o Amazon EC2 Auto Scaling termina instâncias no grupo e executa as substituições para as instâncias terminadas. A política de término para o grupo do Auto Scaling controla quais instâncias são substituídas primeiro.

Rebalanceamento de zona de disponibilidade

O Amazon EC2 Auto Scaling equilibra sua capacidade uniformemente nas zonas de disponibilidade habilitadas para seu grupo do Auto Scaling. Isso ajuda a reduzir o impacto de uma paralisação da zona de disponibilidade. Se a distribuição da capacidade entre zonas de disponibilidade ficar fora de equilíbrio, o Amazon EC2 Auto Scaling reequilibra o grupo do Auto Scaling iniciando instâncias nas zonas de disponibilidade habilitadas com o menor número de instâncias e terminando instâncias em outro lugar. A política de término controla quais instâncias são priorizadas para término primeiro.

Há vários motivos pelos quais a distribuição de instâncias nas zonas de disponibilidade pode ficar fora de equilíbrio.

Remoção de instâncias

Se você desvincular instâncias do seu grupo do Auto Scaling ou terminar instâncias explicitamente e diminuir a capacidade desejada, impedindo assim que as instâncias de substituição sejam executadas, o grupo poderá ficar desbalanceado. Se isso ocorrer, o Amazon EC2 Auto Scaling compensará rebalanceando as zonas de disponibilidade.

Uso de zonas de disponibilidade diferentes das especificadas originalmente

Se você expandir seu grupo do Auto Scaling para incluir zonas de disponibilidade adicionais ou alterar quais zonas de disponibilidade serão usadas, o Amazon EC2 Auto Scaling iniciará instâncias nas novas zonas de disponibilidade e terminará instâncias nas outras zonas para ajudar a garantir que seu grupo do Auto Scaling abranja as zonas de disponibilidade de modo uniforme.

Interrupção de disponibilidade

As interrupções de disponibilidade são raras. No entanto, se uma zona de disponibilidade ficar indisponível e for recuperada posteriormente, seu grupo do Auto Scaling poderá se tornar desbalanceado entre as zonas de disponibilidade. O Amazon EC2 Auto Scaling tenta rebalancear gradualmente o grupo, e o rebalanceamento pode terminar instâncias em outras zonas.

Veja o exemplo em que você tem um grupo do Auto Scaling que tem um tipo de instância, duas zonas de disponibilidade e uma capacidade desejada de duas instâncias. Em uma situação em que uma zona de disponibilidade falha, o Amazon EC2 Auto Scaling executa automaticamente uma nova instância na zona de disponibilidade íntegra para substituir a da zona de disponibilidade não íntegra. Em seguida, quando a zona de disponibilidade não íntegra retorna a um estado íntegro, o Amazon EC2 Auto Scaling executa automaticamente uma nova instância nessa zona, que, por sua vez, termina uma instância na zona não afetada.

> **Note**
>
> No rebalanceamento, o Amazon EC2 Auto Scaling ativa novas instâncias antes de terminar as antigas, para que o processo não comprometa a performance nem a disponibilidade da sua aplicação.
>
> Como o Amazon EC2 Auto Scaling tenta ativar novas instâncias antes de terminar as antigas, estar na capacidade máxima especificada ou próximo a ela pode impedir ou interromper completamente as atividades de rebalanceamento. Para evitar esse problema, o sistema pode exceder temporariamente a capacidade máxima especificada de um grupo em uma margem de 10% (ou em uma margem de uma instância, o que for maior) durante uma atividade de rebalanceamento. A margem é estendida somente se o grupo estiver na capacidade máxima ou próximo a ela e precisar de rebalanceamento, seja devido ao rezoneamento solicitado pelo usuário ou para compensar os problemas de disponibilidade da zona. A extensão dura somente pelo tempo necessário para rebalancear o grupo.

Trabalhar com políticas de término do Amazon EC2 Auto Scaling

Uma política de terminação fornece os critérios que o Amazon EC2 Auto Scaling segue para terminar as instâncias em uma ordem específica. Por padrão, o Amazon EC2 Auto Scaling usa uma política de terminação que foi projetada para terminar primeiro as instâncias que estão usando configurações desatualizadas. Você pode alterar a política de terminação para controlar quais instâncias são mais importantes para serem terminadas primeiro.

Quando o Amazon EC2 Auto Scaling encerra instâncias, ele tenta manter o equilíbrio entre as zonas de disponibilidade que são usadas pelo grupo do Auto Scaling. A manutenção do equilíbrio entre as zonas de disponibilidade tem precedência sobre as políticas de encerramento. Se uma zona de disponibilidade tiver mais instâncias do que outras, o Amazon EC2 Auto Scaling aplicará a política de

terminação à zona desbalanceada primeiro. Se as zonas de disponibilidade estiverem balanceadas, ele aplicará a política de terminação em todas as zonas.

Tópicos

- Como funciona a política de terminação padrão
- Política de término padrão e grupos de instâncias mistas
- Políticas de terminação predefinidas
- Alterar a política de término de um grupo do Auto Scaling

> **Note**
>
> O Amazon EC2 Auto Scaling aplica políticas de rescisão somente às instâncias que não são consideradas insalubres pelo grupo do Auto Scaling. Como resultado, as instâncias marcadas como não íntegras pelas verificações de saúde do Auto Scaling ignorarão a avaliação da política de rescisão.
> Para obter mais informações, consulte Projete suas aplicações para lidar com a terminação de instâncias sem problemas.

Como funciona a política de terminação padrão

Quando o Amazon EC2 Auto Scaling termina instâncias, ele primeiro determina quais zonas de disponibilidade têm a maioria das instâncias, e pelo menos uma instância que não esteja protegida contra redução de escala na horizontal. Em seguida, ele avalia as instâncias desprotegidas dentro da zona de disponibilidade identificada da seguinte forma:

Instâncias que usam configurações desatualizadas

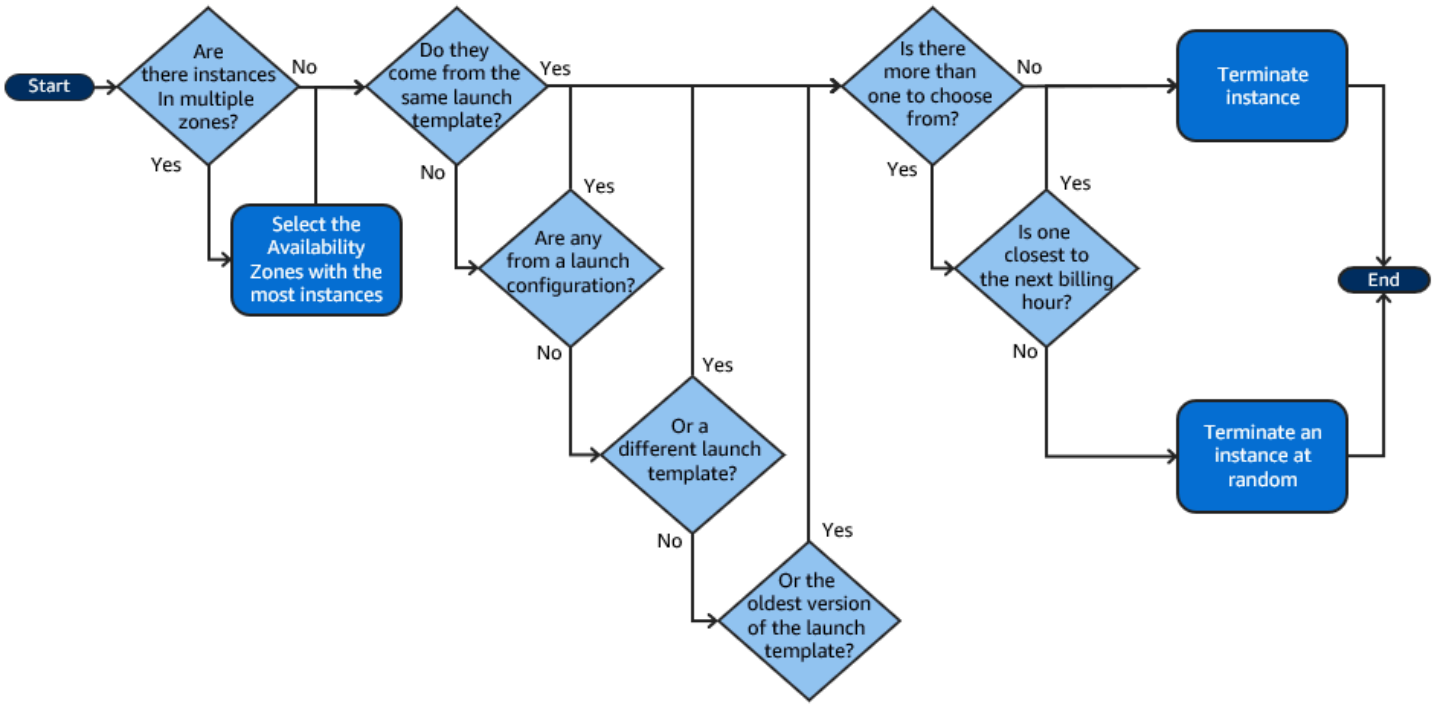
- Para grupos que usam um modelo de execução: Determine se alguma das instâncias usa configurações desatualizadas, priorizando nesta ordem:
  1. Primeiro, verifique se há instâncias executadas com uma configuração de execução.
  2. Em seguida, verifique se há instâncias executadas usando um modelo de execução diferente em vez do modelo de execução atual.
  3. Por fim, verifique se há instâncias usando a versão mais antiga do modelo de execução atual.
- Para grupos que usam uma configuração de execução: determine se qualquer uma das instâncias usa a configuração de ativação mais antiga.

Se não forem encontradas instâncias com configurações desatualizadas ou se houver várias instâncias para escolher, o Amazon EC2 Auto Scaling considerará os próximos critérios de instâncias que se aproximam da próxima hora de faturamento.

Instâncias que se aproximam da próxima hora de faturamento

Determine se alguma das instâncias que atendem aos critérios anteriores está mais próxima da próxima hora de faturamento. Se várias instâncias estiverem igualmente próximas, termine uma delas de forma aleatória. (Isso ajuda a maximizar o uso de suas instâncias que têm uma taxa por hora.) No entanto, a maior parte do uso do EC2 agora é faturada por segundo, portanto, essa otimização oferece menos benefícios. Para obter mais informações, consulte Definição de preço do Amazon EC2.

O diagrama de fluxo a seguir ilustra como a política de terminação padrão funciona para grupos que usam um modelo de execução.



Política de término padrão e grupos de instâncias mistas

O Amazon EC2 Auto Scaling aplica critérios adicionais ao terminar instâncias em grupos de instâncias mistas

Quando o Amazon EC2 Auto Scaling precisa terminar uma instância, ele primeiro identifica qual opção de compra (spot ou sob demanda) deve ser terminada com base nas configurações do grupo.

Isso garante que o grupo tenda a atingir a proporção especificada de instâncias spot e sob demanda ao longo do tempo.

Em seguida, aplica as políticas de término em cada zona de disponibilidade individualmente. Ele determina qual Instância spot ou sob demanda em qual Zona de Disponibilidade deve ser terminada para manter as Zonas de disponibilidade balanceadas. Uma lógica semelhante se aplica a grupos de instâncias mistas com pesos definidos para os tipos de instância.

Em cada zona, a política de terminação padrão funciona da seguinte forma para determinar qual instância desprotegida dentro da opção de compra identificada pode ser terminada:

1. Determine se alguma das instâncias pode ser terminada para melhorar o alinhamento com a estratégia de alocação especificada para o grupo do Auto Scaling. Se nenhuma instância for identificada para otimização ou se houver várias instâncias para escolher, a avaliação continuará.
2. Determine se alguma das instâncias usa configurações desatualizadas, priorizando nesta ordem:
   a. Primeiro, verifique se há instâncias executadas com uma configuração de execução.
   b. Em seguida, verifique se há instâncias executadas usando um modelo de execução diferente em vez do modelo de execução atual.
   c. Por fim, verifique se há instâncias usando a versão mais antiga do modelo de execução atual.

   Se não forem encontradas instâncias com configurações desatualizadas ou se houver várias instâncias para escolher, a avaliação continuará.
3. Determine se alguma das instâncias está mais próxima da próxima hora de faturamento. Se várias instâncias estiverem igualmente próximas, escolha uma delas de forma aleatória.

Políticas de terminação predefinidas

Você escolhe entre as seguintes políticas de terminação predefinidas:

- **`Default`**. Terminar instâncias de acordo com a política de término padrão.
- **`AllocationStrategy`**. Terminar as instâncias no grupo do Auto Scaling para alinhar as instâncias restantes com a estratégia de alocação para o tipo de instância que está sendo terminada (uma instância spot ou uma instância sob demanda). Essa política é útil quando seus tipos de instância preferidos foram alterados. Se a estratégia de alocação spot for `lowest-price`, você poderá rebalancear gradualmente a distribuição de instâncias spot nos seus N grupos spot mais econômicos. Se a estratégia de alocação spot for `capacity-optimized`, você poderá rebalancear gradualmente a distribuição de instâncias spot nos grupos spot onde há mais

capacidade spot disponível. Você também pode substituir gradualmente instâncias sob demanda de um tipo de prioridade mais baixo por instâncias sob demanda de um tipo de prioridade mais alto.

- **`OldestLaunchTemplate`**. Terminar as instâncias que têm o modelo de execução mais antigo. Com essa política, as instâncias que usam o modelo de execução que não é o atual são encerradas primeiro, seguidas pelas instâncias que usam a versão mais antiga do modelo de execução atual. Essa política é útil quando você está atualizando um grupo e descontinuando as instâncias de uma configuração anterior.
- **`OldestLaunchConfiguration`**. Terminar as instâncias que têm a configuração de execução mais antiga. Essa política é útil quando você está atualizando um grupo e descontinuando as instâncias de uma configuração anterior. Com essa política, as instâncias que usem a configuração de execução que não seja a atual são encerradas primeiro.
- **`ClosestToNextInstanceHour`**. Terminar as instâncias que estão mais perto da próxima hora de faturamento. Essa política ajuda a maximizar o uso de suas instâncias que têm uma taxa por hora.
- **`NewestInstance`**. Terminar a instância mais recente do grupo. Essa política é útil quando você está testando uma nova configuração de ativação, mas não deseja mantê-la em produção.
- **`OldestInstance`**. Terminar a instância mais antiga do grupo. Essa opção é útil quando você está atualizando as instâncias no grupo do Auto Scaling para um novo tipo de instância do EC2. Você pode substituir instâncias do tipo antigo gradualmente por instâncias do tipo novo.

> **Note**
>
> O Amazon EC2 Auto Scaling sempre equilibra as instâncias entre as zonas de disponibilidade primeiro, independentemente da política de término usada. Como resultado, você pode encontrar situações em que algumas instâncias mais recentes são terminadas antes de instâncias mais antigas. Por exemplo, quando há uma zona de disponibilidade adicionada mais recentemente ou quando uma zona de disponibilidade tiver mais instâncias que as outras zonas de disponibilidade que sejam usadas pelo grupo.

Alterar a política de término de um grupo do Auto Scaling

Para alterar a política de terminação do seu grupo do Auto Scaling, use um dos métodos a seguir.

Console

Não é possível definir o desaquecimento padrão quando você inicialmente cria um grupo do Auto Scaling no console do Amazon EC2 Auto Scaling. A política de término padrão é usada automaticamente. Depois que o grupo do Auto Scaling tiver sido criado, você poderá substituir a política padrão por uma política de terminação diferente ou por várias políticas de terminação listadas na ordem em que devem ser aplicadas.

Para alterar a política de terminação de um grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Detalhes, escolha Configurações avançadas, Editar.
4. Em Políticas de encerramento, escolha uma ou mais políticas de encerramento. Se escolher várias políticas, coloque-as na ordem em que você deseja que elas sejam avaliadas.

   Você tem a opção de escolher Custom termination policy (Política personalizada de encerramento) e, em seguida, escolhe uma função Lambda que atenda às suas necessidades. Se tiver criado versões e aliases para sua função Lambda, é possível escolher uma versão ou alias no menu suspenso Version/Alias (Versão/alias). Para usar a versão não publicada da sua função Lambda, mantenha Version/Alias (Versão/alias) definido como padrão. Para obter mais informações, consulte Criar uma política de término personalizada com o Lambda.

   > **Note**
   >
   > Ao usar várias políticas, a ordem delas devem ser definida corretamente:
   >
   > - Se você usar a política Default (Padrão), coloque-a em último lugar na lista.
   > - Se você usar uma Custom termination policy (Política personalizada de encerramento), ela deve ser a primeira política na lista.
5. Selecione Atualizar.

AWS CLI

A política de término padrão é usada automaticamente, a menos que uma política diferente seja especificada.

Para alterar a política de terminação de um grupo do Auto Scaling

Use um dos seguintes comandos:

- create-auto-scaling-group
- update-auto-scaling-group

Você pode usar as políticas de término individualmente ou combiná-las em uma lista de políticas. Por exemplo, use o comando a seguir para atualizar um grupo do Auto Scaling a fim de usar primeiro a política `OldestLaunchConfiguration` e, depois disso, usar a política `ClosestToNextInstanceHour`.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --
termination-policies "OldestLaunchConfiguration" "ClosestToNextInstanceHour"
```

Se você usar a política de encerramento `Default`, coloque-a no final da lista de políticas de encerramento. Por exemplo, `.--termination-policies "OldestLaunchConfiguration" "Default"`

Para usar uma política de rescisão personalizada, você deve primeiro criar sua política de rescisão usando AWS Lambda. Para especificar a função do Lambda a ser usada como política de término, torne-a a primeira na lista de políticas de término. Por exemplo, `.--termination-policies "arn:aws:lambda:us-west-2:123456789012:function:HelloFunction:prod" "OldestLaunchConfiguration"` Para obter mais informações, consulte Criar uma política de término personalizada com o Lambda.

Criar uma política de término personalizada com o Lambda

O Amazon EC2 Auto Scaling usa políticas de término para priorizar quais instâncias serão terminadas primeiro ao diminuir o tamanho do seu grupo do Auto Scaling (referido como redução de escala na horizontal). O grupo do Auto Scaling usa uma política de término padrão, mas você pode, opcionalmente, escolher ou criar suas próprias políticas de término. Para obter mais informações

sobre como escolher uma política de término predefinida, consulte Trabalhar com políticas de término do Amazon EC2 Auto Scaling.

Neste tópico, você aprenderá como criar uma política de término personalizada usando uma função do AWS Lambda que o Amazon EC2 Auto Scaling chama em resposta a determinados eventos. A função do Lambda que você cria processa as informações nos dados de entrada enviados pelo Amazon EC2 Auto Scaling e devolve uma lista de instâncias que estão prontas para término.

Uma política de término personalizada fornece melhor controle sobre quais instâncias são terminadas e quando. Por exemplo, quando seu grupo do Auto Scaling sofre redução de escala na horizontal, o Amazon EC2 Auto Scaling não pode determinar se há workloads em execução que não devem ser interrompidas. Com uma função do Lambda, você pode validar a solicitação de término e aguardar até que a workload seja concluída antes de retornar o ID da instância ao Amazon EC2 Auto Scaling para término.

Conteúdo

- Dados de entrada
- Dados de resposta
- Considerações
- Criar a função do Lambda
- Limitações

Dados de entrada

O Amazon EC2 Auto Scaling gera uma carga útil JSON para eventos de redução de escala na horizontal e também faz isso quando as instâncias estão prestes a ser terminadas como resultado da duração máxima da instância ou dos recursos de atualização da instância. Ele também gera uma carga JSON para os eventos de redução de escala na horizontal que podem ser iniciados ao rebalancear seu grupo nas zonas de disponibilidade.

Essa carga contém informações sobre a capacidade que o Amazon EC2 Auto Scaling precisa terminar, uma lista de instâncias sugeridas para término e o evento que iniciou o término.

Esta é uma carga útil de exemplo:

```
{
  "AutoScalingGroupARN": "arn:aws:autoscaling:us-east-1:<account-
id>:autoScalingGroup:d4738357-2d40-4038-ae7e-b00ae0227003:autoScalingGroupName/my-asg",
```

```
  "AutoScalingGroupName": "my-asg",
  "CapacityToTerminate": [
    {
      "AvailabilityZone": "us-east-1b",
      "Capacity": 2,
      "InstanceMarketOption": "on-demand"
    },
    {
      "AvailabilityZone": "us-east-1b",
      "Capacity": 1,
      "InstanceMarketOption": "spot"
    },
    {
      "AvailabilityZone": "us-east-1c",
      "Capacity": 3,
      "InstanceMarketOption": "on-demand"
    }
  ],
  "Instances": [
    {
      "AvailabilityZone": "us-east-1b",
      "InstanceId": "i-0056faf8da3e1f75d",
      "InstanceType": "t2.nano",
      "InstanceMarketOption": "on-demand"
    },
    {
      "AvailabilityZone": "us-east-1c",
      "InstanceId": "i-02e1c69383a3ed501",
      "InstanceType": "t2.nano",
      "InstanceMarketOption": "on-demand"
    },
    {
      "AvailabilityZone": "us-east-1c",
      "InstanceId": "i-036bc44b6092c01c7",
      "InstanceType": "t2.nano",
      "InstanceMarketOption": "on-demand"
    },
    ...
  ],
  "Cause": "SCALE_IN"
}
```

A carga útil inclui o nome do grupo do Auto Scaling, seu nome do recurso da Amazon (ARN) e os seguintes elementos:

- `CapacityToTerminate` descreve o quanto da sua capacidade spot ou sob demanda está definida para ser terminada em uma determinada zona de disponibilidade.
- `Instances` representa as instâncias que o Amazon EC2 Auto Scaling sugere para término com base nas informações em `CapacityToTerminate`.
- `Cause` descreve o evento que acionou o término `SCALE_IN`, `INSTANCE_REFRESH`, `MAX_INSTANCE_LIFETIME` ou `REBALANCE`.

As informações a seguir descrevem os fatores mais significativos em como o Amazon EC2 Auto Scaling gera as `Instances` nos dados de entrada:

- A manutenção do equilíbrio entre as zonas de disponibilidade tem precedência quando uma instância está sendo terminada devido a eventos de aumento de escala na horizontal e términos baseados na atualização de instância. Dessa forma, se uma zona de disponibilidade tiver mais instâncias que as outras que são usadas pelo grupo, os dados de entrada contêm instâncias qualificáveis para término somente a partir da zona de disponibilidade desbalanceada. Se as zonas de disponibilidade usadas pelo grupo forem balanceadas, os dados de entrada conterão instâncias de todas as zonas de disponibilidade do grupo.
- Ao usar uma política de instâncias mistas, a manutenção das suas capacidades spot e sob demanda em equilíbrio com base nos percentuais desejados para cada opção de compra também tem precedência. Primeiro, identificamos qual dos dois tipos (spot ou sob demanda) deve ser terminado. Em seguida, identificamos quais instâncias (dentro da opção de compra identificada) em que zonas de disponibilidade serão terminadas que resultarão no maior equilíbrio das zonas de disponibilidade.

Dados de resposta

Os dados de entrada e os dados de resposta trabalham juntos para restringir a lista de instâncias a serem terminadas.

Com a entrada dada, a resposta de sua função do Lambda deve se parecer com o exemplo a seguir:

```
{
  "InstanceIDs": [
    "i-02e1c69383a3ed501",
```

```
    "i-036bc44b6092c01c7",
    ...
  ]
}
```

Os `InstanceIDs` na resposta representam as instâncias que estão prontas para sere terminadas.

Como alternativa, você pode devolver um conjunto diferente de instâncias que estão prontas para serem terminadas, o que substitui as instâncias nos dados de entrada. Se nenhuma instância estiver pronta para ser terminada quando sua função do Lambda for chamada, você também pode optar por não devolver nenhuma instância.

Quando não houver nenhuma instância pronta para encerramento, a resposta de sua função do Lambda deverá se parecer com o exemplo a seguir:

```
{
  "InstanceIDs": [ ]
}
```

Considerações

Observe as seguintes considerações ao usar uma política de término personalizada:

- Devolver uma instância primeiro nos dados de resposta não garante seu término. Se mais do que o número necessário de instâncias for devolvido quando sua função do Lambda for chamada, o Amazon EC2 Auto Scaling avaliará cada instância em relação às outras políticas de término especificadas para seu grupo do Auto Scaling. Quando há várias diretivas de término, ele tenta aplicar a próxima diretiva de término na lista e, se houver mais instâncias do que as necessárias para término, ele passa para a próxima diretiva de término, e assim por diante. Se nenhuma outra política de término for especificada, a política de término padrão será usada para determinar quais instâncias serão terminadas.
- Se nenhuma instância for devolvida ou se sua função do Lambda expirar, o Amazon EC2 Auto Scaling aguardará um curto período de tempo antes de chamar sua função novamente. Para qualquer evento de redução de escala na horizontal, ele continua tentando desde que a capacidade desejada do grupo seja menor que sua capacidade atual. Por exemplo, términos baseados em atualização, ele continua tentando por uma hora. Depois disso, se continuar a falhar ao terminar quaisquer instâncias, a operação de atualização da instância falhará. Com a duração máxima da instância, o Amazon EC2 Auto Scaling continua tentando terminar a instância identificada como excedendo sua vida útil máxima.

- Como sua função é repetida continuamente, certifique-se de testar e corrigir quaisquer erros permanentes em seu código antes de usar uma função do Lambda como uma política de término personalizada.
- Se você substituir os dados de entrada com sua própria lista de instâncias a serem terminadas, e o término dessas instâncias deixar as zonas de disponibilidade fora de equilíbrio, o Amazon EC2 Auto Scaling reequilibrará gradualmente a distribuição de capacidade entre zonas de disponibilidade. Primeiro, ele invoca sua função do Lambda para ver se existem instâncias que estão prontas para serem terminadas para que ele possa determinar se deseja iniciar o rebalanceamento. Se houver instâncias prontas para serem terminadas, ele iniciará novas instâncias primeiro. Quando as instâncias terminam de ser iniciadas, elas detectam que a capacidade atual do grupo é maior do que a capacidade desejada e iniciam um evento de redução de escala na horizontal.
- Uma política de encerramento personalizada não afeta sua capacidade de também usar proteção para reduzir a escala horizontalmente para evitar que determinadas instâncias sejam encerradas. Para obter mais informações, consulte Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância.
- Se sua função Lambda não retornar nenhuma instância, isso não impedirá automaticamente todos os encerramentos. O Amazon EC2 Auto Scaling ainda encerrará instâncias consideradas insalubres pelas verificações de saúde do Auto Scaling, independentemente das políticas de rescisão.

Criar a função do Lambda

Comece criando a função do Lambda, para que você possa especificar seu nome do recurso da Amazon (ARN) nas políticas de término do seu grupo do Auto Scaling.

Para criar uma função do Lambda (console)

1. Abra a página Functions (Funções) no console do Lambda.
2. Na barra de navegação na parte superior da tela, escolha a mesma região usada ao criar o grupo do Auto Scaling.
3. Escolha Create function (Criar função) e Author from scratch (Criar desde o início).
4. Em Basic information (Informações básicas), para Function name (Nome da função), insira um nome para a função.
5. Escolha a opção Criar função. Você é retornado ao código e configuração da função.

6. Com sua função ainda aberta no console, em Function code (Código da função), cole seu código no editor.
7. Escolha Implantar.
8. Opcionalmente, crie uma versão publicada da função do Lambda escolhendo a guia Versions (Versões), e depois, Publish new version (Publicar nova versão). Para saber mais sobre controle de versões no Lambda, consulte Versões de função do Lambda no Guia do desenvolvedor do AWS Lambda .
9. Se você optou por publicar uma versão, escolha a guia Aliases caso deseje associar um alias a essa versão da função do Lambda. Para saber mais sobre aliases no Lambda, consulte Aliases de função do Lambda no Guia do desenvolvedor do AWS Lambda .
10. Em seguida, escolha a guia Configuration (Configuração) e, depois, Permissions (Permissões).
11. Role para baixo até Resource-based policy (Política baseada em recurso) e, em seguida, escolha Add permissions (Adicionar permissões). Uma política baseada em recurso é usada para conceder permissões para invocar sua função no principal que é especificado na política. Neste caso, o principal será o perfil vinculado ao serviço do Amazon EC2 Auto Scaling que está associada ao grupo do Auto Scaling.
12. Na seção Policy statement (Declaração da política), configure suas permissões:
    a. Selecione Conta da AWS.
    b. Em Principal insira o ARN da função vinculada a serviço de chamada, por exemplo, **`arn:aws:iam::<aws-account-id>:role/aws-service-role/autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling`**.
    c. Para Ação, escolha lambda: InvokeFunction.
    d. Em Statement ID (ID da instrução), insira um ID de instrução exclusivo, como **`AllowInvokeByAutoScaling`**.
    e. Escolha Salvar.
13. Depois que você tiver seguido essas instruções, prossiga para especificar o ARN de sua função nas políticas de término do grupo do Auto Scaling como próxima etapa. Para obter mais informações, consulte Alterar a política de término de um grupo do Auto Scaling.

> **Note**
>
> Para exemplos que você pode usar como referência para desenvolver sua função Lambda, consulte o GitHub repositório do Amazon EC2 Auto Scaling.

Limitações

- Você só pode especificar uma função do Lambda nas políticas de término para um grupo do Auto Scaling. Se houver várias políticas de término especificadas, a função do Lambda deve ser especificada primeiro.
- Você pode referenciar sua função do Lambda usando um ARN não qualificado (sem um sufixo) ou um ARN qualificado que tenha uma versão ou um alias como sufixo. Se um ARN não qualificado for usado (por exemplo, `function:my-function`), sua política baseada em recurso deve ser criada na versão não publicada da sua função. Se um ARN qualificado for usado (por exemplo, `function:my-function:1` ou `function:my-function:prod`), sua política baseada em recurso deve ser criada na versão publicada específica da sua função.
- Você não pode usar um ARN qualificado com o sufixo `$LATEST`. Se você tentar adicionar uma política de término personalizada que se refira a um ARN qualificado com o sufixo `$LATEST`, isso resultará em um erro.
- O número de instâncias fornecidas nos dados de entrada é limitado a 30.000 instâncias. Se houver mais de 30.000 instâncias que possam ser terminadas, os dados de entrada incluirão `"HasMoreInstances": true` para indicar que o número máximo de instâncias é devolvido.
- O tempo máximo de execução para sua função do Lambda é de dois segundos (2000 milissegundos). Como prática recomendada, é necessário definir o valor de tempo-limite da função do Lambda com base no tempo de execução esperado. As funções do Lambda têm um tempo limite padrão de três segundos, mas isso pode ser reduzido.
- Se o runtime exceder o limite de 2 segundos, qualquer ação de redução da escala horizontalmente ficará em espera até que o tempo de execução caia abaixo desse limite. No caso de funções de Lambda com runtimes consistentemente mais longos, encontre uma maneira de reduzir o runtime, por exemplo, armazenando em cache os resultados onde eles possam ser recuperados durante as invocações de Lambda subsequentes.
- Não há suporte a funções do Lambda entre contas. A função Lambda usada como política de encerramento personalizada deve estar na Conta da AWS mesma do grupo Auto Scaling. As instruções de configuração nesta página se concentram especificamente nas configurações da mesma conta.
- As políticas de encerramento personalizadas não se aplicam a instâncias não íntegras. As instâncias não íntegras são sempre consideradas elegíveis para rescisão, e a função Lambda avalia apenas a ordem de rescisão das instâncias restantes.

Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância

A proteção de redução da escala horizontalmente da instância lhe dá controle sobre quais instâncias o Amazon EC2 Auto Scaling pode terminar. Um caso de uso comum para esse requisito é escalar workload baseados em contêineres. Para obter mais informações, consulte Projete suas aplicações para lidar com a terminação de instâncias sem problemas.

Por padrão, a proteção de redução de escala horizontalmente da instância é desativada quando você cria um grupo do Auto Scaling Isso significa que o Amazon EC2 Auto Scaling pode terminar qualquer instância no grupo.

É possível proteger as instâncias assim que elas são iniciadas ao habilitar a configuração de proteção contra redução de escala na horizontal de instâncias no seu grupo do Auto Scaling. A proteção de redução de instâncias começa quando o estado da instância é `InService`. Em seguida, para controlar quais instâncias podem ser encerradas, desabilite a configuração de proteção escalável em instâncias individuais dentro do grupo do Auto Scaling. Ao fazer isso, você pode continuar protegendo determinadas instâncias contra encerramentos indesejados.

Tópicos

- Considerações
- Alterar a proteção de redução de escala horizontalmente de um grupo do Auto Scaling
- Alterar a proteção de redução de escala horizontalmente de uma instância

Considerações

Veja a seguir algumas considerações ao usar a proteção de redução de escala horizontalmente de instância:

- Se todas as instâncias de um grupo do Auto Scaling estiverem protegidas contra a redução de escala na horizontal e ocorrer um evento de redução de escala na horizontal, a capacidade desejada será reduzida. No entanto, o grupo do Auto Scaling não pode terminar o número necessário de instâncias até que suas configurações de proteção contra redução de escala na horizontal de instâncias sejam desabilitadas. No Console de gerenciamento da AWS, o histórico de atividades do grupo Auto Scaling inclui a seguinte mensagem se todas as instâncias em um grupo de Auto Scaling estiverem protegidas da escalabilidade quando ocorrer um evento de

escalabilidade: `Could not scale to desired capacity because all remaining instances are protected from scale in.`

- Se você desvincular uma instância protegida contra redução de escala na horizontal, sua configuração de proteção de redução de instâncias será perdida. Quando a instância é associada ao grupo novamente, ela herda a configuração de proteção de redução de instâncias atual do grupo. Quando o Amazon EC2 Auto Scaling executa uma instância ou move uma instância de um pool ativo para um grupo do Auto Scaling, a instância herda a configuração de proteção contra redução da escala de instâncias na horizontal do grupo do Auto Scaling.
- A proteção contra redução de escala na horizontal de instâncias não protege as instâncias do Auto Scaling contra o seguinte:
  - Substituição da verificação se a instância não passar nas verificações de integridade. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.
  - Interrupções de instâncias spot Uma instância spot é encerrada quando a capacidade não está mais disponível ou o preço spot excede seu preço máximo.
  - Reserva de blocos de capacidade O Amazon EC2 recupera as instâncias do bloco de capacidade, mesmo que elas estejam protegidas contra redução de escala horizontalmente.
  - Terminação manual por meio do comando `terminate-instance-in-auto-scaling-group`. Para obter mais informações, consulte Encerrar uma instância no seu grupo do Auto Scaling (AWS CLI).
  - Terminação manual por meio do console do Amazon EC2, comandos da CLI e operações de API. Para proteger as instâncias do Auto Scaling contra término manual, habilite a proteção contra término do Amazon EC2. (Isso não impede que o Amazon EC2 Auto Scaling termine as instâncias ou a terminação manual por meio do comando `terminate-instance-in-auto-scaling-group`). Para obter informações sobre como ativar a proteção de terminação do Amazon EC2 em um modelo de execução, consulte Criar um modelo de execução usando configurações avançadas.

Alterar a proteção de redução de escala horizontalmente de um grupo do Auto Scaling

É possível habilitar ou desabilitar a configuração de proteção contra redução de escala na horizontal de instâncias para um grupo do Auto Scaling. Quando você a ativa, todas as novas instâncias iniciadas pelo grupo terão a proteção de redução de escala horizontalmente de instância ativada.

A ativação ou desativação dessa configuração para um grupo do Auto Scaling não afeta as instâncias existentes.

Console

Para ativar a proteção de redução de escala horizontalmente de um novo grupo do Auto Scaling

Ao criar o grupo do Auto Scaling, na página Configure group size and scaling policies (Configurar tamanho do grupo e políticas de escalabilidade), em Instance scale-in protection (Proteção contra redução de escala na horizontal de instâncias), selecione a opção Enable instance scale-in protection (Habilitar proteção contra redução de escala na horizontal de instâncias).

Para ativar ou desativar a proteção de redução de escala horizontalmente para um grupo existente

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Detalhes, escolha Configurações avançadas, Editar.
4. Para Proteção de redução de escala horizontalmente de instância, marque ou desmarque a caixa de seleção Ativar proteção de redução de escala horizontalmente de instância para ativar ou desativar essa opção conforme necessário.
5. Selecione Atualizar.

AWS CLI

Para ativar a proteção de redução de escala horizontalmente de um novo grupo do Auto Scaling

Use o seguinte comando create-auto-scaling-group para habilitar a proteção de redução de instâncias.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg --new-
instances-protected-from-scale-in ...
```

Ativar a proteção contra redução de escala na horizontal de instâncias para um grupo

Use o update-auto-scaling-groupcomando a seguir para ativar a proteção de escalabilidade da instância para o grupo de Auto Scaling especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --new-
instances-protected-from-scale-in
```

Para desativar a proteção de redução de escala horizontalmente para um grupo existente

Use o seguinte comando para desabilitar a proteção de redução de instâncias para o grupo especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --no-new-
instances-protected-from-scale-in
```

Alterar a proteção de redução de escala horizontalmente de uma instância

Por padrão, uma instância obtém sua configuração de proteção contra redução de escala na horizontal de instâncias de seu grupo do Auto Scaling. No entanto, é possível habilitar ou desabilitar a proteção de redução para uma instância a qualquer momento.

Console

Para ativar ou desativar a proteção de redução de escala horizontalmente para uma instância

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), selecione uma instância.
4. Para habilitar a proteção de redução de instâncias, escolha Actions (Ações) e Set scale-in protection (Definir proteção de redução). Quando solicitado, escolha Set scale-in protection (Definir proteção de redução).
5. Para desabilitar a proteção de redução de instâncias, escolha Actions (Ações) e Remove scale-in protection (Remover proteção de redução). Quando solicitado, escolha Remove Scale In Protection (Remover proteção de redução).

AWS CLI

Como habilitar a proteção de redução de instâncias ()

Use o seguinte comando set-instance-protection para habilitar a proteção de redução para a instância especificada.

```
aws autoscaling set-instance-protection --instance-ids i-5f2e8a0d --auto-scaling-
group-name my-asg --protected-from-scale-in
```

Para desativar a proteção de redução de escala horizontalmente para uma instância

Use o seguinte comando para desabilitar a proteção de redução para a instância especificada,

```
aws autoscaling set-instance-protection --instance-ids i-5f2e8a0d --auto-scaling-
group-name my-asg --no-protected-from-scale-in
```

> **Note**
>
> A proteção de redução de escala na horizontal de instâncias não garante que as instâncias não serão terminadas no caso de um erro humano, como, por exemplo, se alguém terminar manualmente uma instância usando o console do Amazon EC2 ou a AWS CLI. Para proteger sua instância contra término acidental, use a proteção contra término do Amazon EC2. No entanto, mesmo com a proteção contra término e a proteção de aumento de escala na horizontal de instâncias habilitadas, os dados salvos no armazenamento da instância podem ser perdidos se uma verificação de integridade determinar que uma instância não está íntegra ou se o próprio grupo for excluído acidentalmente. Como em qualquer ambiente, uma prática recomendada é fazer backup de seus dados com frequência ou sempre que for apropriado para seus requisitos de continuidade de negócios.

Projete suas aplicações para lidar com a terminação de instâncias sem problemas

Este tópico descreve os recursos que você pode usar para impedir que seu grupo do Amazon EC2 Auto Scaling encerre instâncias do Amazon EC2 que ainda não estão prontas para serem encerradas. Por padrão, o Auto Scaling não tem visibilidade dos aplicativos em execução nas suas instâncias. Ele pode encerrar instâncias antes que seu aplicativo possa ser encerrado ou

concluído normalmente as tarefas atribuídas. Esses recursos dão ao seu aplicativo tempo para concluir o trabalho em andamento, transferir o estado ou realizar a limpeza antes do encerramento da instância. Você pode usá-los individualmente ou em combinação, dependendo dos requisitos do seu aplicativo.

Esses recursos são particularmente úteis para cargas de trabalho com estado, em que cada instância em sua frota contém dados, trabalhos ou estados diferentes de outras instâncias. O encerramento de instâncias com estado sem um desligamento normal pode resultar na reinicialização de trabalhos de longa duração desde o início, na redução da redundância ou perda de dados e na interrupção de transações ou cálculos em andamento. Para encerrar normalmente uma instância com estado, sua carga de trabalho deve ser drenada (concluindo todas as tarefas atualmente atribuídas) ou transferida (movendo tarefas, dados ou configuração para outra instância ativa).

Conteúdo

- Ganchos do ciclo de vida de encerramento
- Proteção de redução horizontal de escala de instâncias
- Política de encerramento personalizada
- Política de ciclo de vida da instância
- Suspender completamente as rescisões
- Limitações
- Cenários de exemplo

Ganchos do ciclo de vida de encerramento

Um gancho do ciclo de vida de encerramento prolonga a vida útil da sua instância do Amazon EC2 que já está selecionada para encerramento. Ele fornece tempo extra para concluir o trabalho em andamento atualmente atribuído à instância ou para salvar o progresso e transferir o trabalho para outra instância.

Para muitas cargas de trabalho, um gancho do ciclo de vida de encerramento pode ser suficiente para encerrar normalmente um aplicativo em uma instância selecionada para encerramento. Essa é a melhor abordagem e não pode ser usada para evitar a rescisão se uma ação do ciclo de vida da rescisão for abandonada. Se sua carga de trabalho tiver uma baixa tolerância a falhas na ação do ciclo de vida de encerramento, configure as políticas de ciclo de vida da instância em combinação com seus ganchos do ciclo de vida de encerramento para reter instâncias.

Para usar um gancho de ciclo de vida de encerramento, você precisa saber quando uma instância é selecionada para encerramento. Você tem duas maneiras de saber isso:

Opção	Description	Melhor usado para	Link para a documentação
No interior da instância	O serviço de metadados de instância (IMDS) é um endpoint seguro que você pode pesquisar o status de uma instância diretamente da instância. Se os metadados voltarem com,`Terminated` sua instância está programada para ser encerrada.	Aplicações em que é necessário realizar uma ação na instância antes que ela seja encerrada.	Recupere o estado do ciclo de vida de destino
Fora da instância	Quando uma instância é encerrada, uma notificação de evento é gerada. Você pode criar regras usando Amazon EventBridge, Amazon SQS, Amazon SNS AWS Lambda ou capturar esses eventos e invocar uma resposta, como com uma função Lambda.	Aplicações que precisam agir fora da instância.	Configurar um alvo de notificação

Para usar um gancho do ciclo de vida, você também precisa saber quando sua instância está pronta para ser totalmente encerrada. O Amazon EC2 Auto Scaling não encerrará a instância até que ela CompleteLifecycleActionreceba uma chamada de API ou que o tempo limite termine, o que ocorrer primeiro.

Por padrão, uma instância pode continuar em execução por uma hora (tempo limite de pulsação) devido a um gancho do ciclo de vida de encerramento. Você pode configurar o tempo limite padrão se uma hora não for suficiente para concluir a ação do ciclo de vida. Quando uma ação do ciclo de vida está em andamento, você pode estender o tempo limite com RecordLifecycleActionHeartbeatchamadas de API.

Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

Proteção de redução horizontal de escala de instâncias

Você pode usar a proteção de escalabilidade de instâncias para controlar quais instâncias são selecionadas para encerramento durante eventos de expansão, especialmente para evitar que uma instância que esteja processando ativamente um trabalho de longa duração seja encerrada. Por exemplo, ao executar cargas de trabalho em contêineres, é comum querer proteger todas as instâncias e remover a proteção somente para instâncias sem tarefas atuais ou agendadas. As instâncias podem continuar pesquisando novos trabalhos e reativar a proteção quando houver novos trabalhos atribuídos.

Você pode ativar a proteção escalável no nível do grupo e da instância do Auto Scaling. Quando você ativa a proteção escalável no nível do grupo Auto Scaling, somente novas instâncias são protegidas quando são criadas. Para instâncias existentes, você pode ativar a proteção individualmente.

Os aplicativos podem definir a proteção a partir das próprias instâncias ou de um plano de controle centralizado que gerencia se cada instância pode ser encerrada. Recomendamos a abordagem centralizada para grandes frotas ou quando a proteção precisa ser alternada com frequência, pois ela permite que você faça chamadas em lote SetInstanceProtectione evite problemas de limitação de API.

Para obter mais informações, consulte Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância.

Política de encerramento personalizada

Assim como a proteção escalável de instâncias, uma política de rescisão personalizada ajuda a evitar que seu grupo Amazon EC2 Auto Scaling encerre instâncias específicas do EC2. As instâncias não íntegras ainda podem ser encerradas, independentemente da sua política de encerramento personalizada.

Seu grupo de Auto Scaling usa uma política de encerramento padrão para determinar quais instâncias do Amazon EC2 ele encerra primeiro. Se você quiser ter mais controle sobre quais instâncias serão encerradas primeiro, você pode implementar uma política de encerramento personalizada usando uma função Lambda. O Auto Scaling chama essa função sempre que precisar selecionar uma instância para encerramento e encerrará somente as instâncias retornadas pela função. Se a função errar, atingir o tempo limite ou retornar uma lista vazia, o Auto Scaling não encerrará nenhuma instância, a menos que ela não esteja íntegra.

Uma política de encerramento personalizada é útil quando seu aplicativo pode identificar quais instâncias estão ociosas ou seguras de encerrar. Isso normalmente requer um plano de controle que rastreie a carga de trabalho em todo o grupo.

Para obter mais informações, consulte Criar uma política de término personalizada com o Lambda.

Política de ciclo de vida da instância

As políticas de ciclo de vida da instância oferecem proteção contra rescisões do Amazon EC2 Auto Scaling quando uma ação do ciclo de vida de rescisão é abandonada. Diferentemente dos ganchos de ciclo de vida isolados, as políticas de ciclo de vida das instâncias são projetadas para garantir que as instâncias sejam transferidas para um estado retido quando procedimentos de desligamento regulares não forem concluídos com êxito.

Quando o Auto Scaling seleciona uma instância para encerramento, seus ganchos de ciclo de vida de terminação configurados são invocados e seu aplicativo inicia procedimentos de desligamento adequados. Se as ações do ciclo de vida de encerramento forem concluídas com êxito`CONTINUE`, a instância será encerrada normalmente. No entanto, se uma ação de encerramento do ciclo de vida for abandonada por qualquer motivo, a política de ciclo de vida da instância moverá a instância para um estado retido em vez de encerrá-la. As instâncias retidas não contam para a capacidade desejada pelo seu grupo de Auto Scaling, então as instâncias substitutas são iniciadas automaticamente. Você incorrerá em cobranças padrão do Amazon EC2 tanto pela instância retida quanto por sua substituição até que você encerre manualmente a instância retida usando a API. TerminateInstanceInAutoScalingGroup

Para usar esse recurso, você deve configurar uma política de ciclo de vida da instância com o gatilho de `TerminateHookAbandon` retenção definido como`retain`, bem como pelo menos um gancho de ciclo de vida de encerramento. Como as instâncias retidas incorrem em custos contínuos do Amazon EC2 e exigem ação manual, o monitoramento é fundamental. Você deve ativar CloudWatch métricas como `GroupTerminatingRetainedInstances` e criar CloudWatch alarmes para alertá-lo quando as instâncias entrarem em estados retidos.

Para obter mais informações, consulte Controle a retenção de instâncias com políticas de ciclo de vida de instâncias.

Suspender completamente as rescisões

Se você precisar de controle total sobre todas as terminações de instâncias do EC2 dentro do seu grupo Amazon EC2 Auto Scaling, suspenda o processo. `Terminate` Só recomendamos usar

essa opção se as opções acima não oferecerem o controle necessário para o seu serviço. Ao ligar SuspendProcessespara suspender o `Terminate` processo, você evita que o Auto Scaling tente encerramentos por qualquer motivo, exceto aqueles iniciados por uma solicitação do usuário à API. TerminateInstanceInAutoScalingGroup

Para obter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.

Limitações

> **Important**
>
> Ao projetar seu aplicativo no Amazon EC2 Auto Scaling para lidar com terminações de instâncias sem problemas, tenha em mente as seguintes limitações.

Instâncias insalubres ignoram algumas proteções

Se uma instância não estiver íntegra, o Amazon EC2 Auto Scaling começará a encerrá-la mesmo que você tenha políticas de encerramento personalizadas ou proteção de escalabilidade em vigor. A única maneira de evitar a substituição de instâncias não íntegras pelo Auto Scaling é suspender `HealthCheck` o processo`ReplaceUnhealthy`,, ou. `Terminate` Você pode usar ganchos de ciclo de vida e uma política de ciclo de vida da instância para permitir que o aplicativo seja desligado normalmente ou copie quaisquer dados que você precise recuperar antes que a instância não íntegra seja encerrada.

Recurso	Controla instâncias íntegras	Controla instâncias não íntegras
Políticas de rescisão personalizadas	Sim	Não
Proteção contra redução	Sim	Não
Suspender `HealthCheck` `ReplaceUnhealthy` ou processar `Terminate`	Yes (Sim)	Yes (Sim)
Hooks do ciclo de vida	Yes (Sim)	Yes (Sim)
Política de ciclo de vida da instância	Yes (Sim)	Yes (Sim)

Os ganchos de ciclo de vida por si só não garantem um desligamento normal

Por padrão, os ganchos do ciclo de vida da terminação operam com base no melhor esforço. Se uma ação do ciclo de vida de encerramento for abandonada, o Amazon EC2 Auto Scaling prosseguirá com o encerramento imediato da instância. Você pode combinar ganchos do ciclo de vida de encerramento com uma política de ciclo de vida de instâncias para reter instâncias quando as ações do ciclo de vida de encerramento forem abandonadas. Com essa combinação:

- Seus ganchos do ciclo de vida de encerramento tentam desligar seu aplicativo normalmente após o Auto Scaling acionar o encerramento da instância e a drenagem de qualquer load balancer configurado do Elastic Load Balancing ser concluída.
- Se uma ação do ciclo de vida de encerramento for abandonada por qualquer motivo, a instância passa para um estado retido em vez de ser encerrada.
- A instância retida permanece em seu estado atual do Amazon EC2, permitindo que você conclua manualmente seus procedimentos de desligamento ou investigue a falha.
- Você pode encerrar manualmente as instâncias retidas chamando a TerminateInstanceInAutoScalingGroupAPI depois de concluir as ações necessárias.

Para obter mais informações, consulte Controle a retenção de instâncias com políticas de ciclo de vida de instâncias.

Certas opções de mercado podem ser interrompidas com aviso limitado

Se você usa opções de mercado de instâncias, como instâncias spot e reservas de capacidade interruptível em seu grupo de Auto Scaling, o Amazon EC2 pode interromper e recuperar suas instâncias a qualquer momento. Essas interrupções ignoram todos os mecanismos de proteção do Amazon EC2 Auto Scaling, incluindo:

- Ganchos do ciclo de vida de encerramento
- Proteção de redução horizontal de escala de instâncias
- Políticas de rescisão personalizadas
- Políticas de ciclo de vida da instância
- Processos suspensos

Quando uma Instância Spot recebe um aviso de interrupção, você tem aproximadamente dois minutos para realizar tarefas de desligamento sem problemas. Embora você possa usar ganchos

do ciclo de vida de encerramento para responder às interrupções da Instância Spot, a instância será encerrada à força no final da janela de dois minutos, mesmo que o gancho do ciclo de vida ainda esteja em andamento. As políticas de ciclo de vida da instância também não podem evitar interrupções na Instância Spot.

Para obter mais informações sobre como lidar com interrupções de instâncias spot, consulte Interrupções de instâncias spot e melhores práticas para Amazon EC2 Spot no Guia do usuário do Amazon EC2.

As terminações diretas do Amazon EC2 ignoram todas as proteções

Se você encerrar uma instância em seu grupo de Auto Scaling diretamente com a API do Amazon TerminateInstancesEC2, o encerramento ignorará todos os mecanismos de proteção do Amazon EC2 Auto Scaling.

Para encerrar instâncias em seu grupo de Auto Scaling respeitando suas proteções configuradas, use a API em vez disso. TerminateInstanceInAutoScalingGroup

Cenários de exemplo

Ao usar o Amazon EC2 Auto Scaling, você pode escolher quanto gerenciamento de frota o Auto Scaling gerencia em seu nome versus quanto controle direto você mantém sobre as decisões de encerramento de instâncias do EC2. Quanto mais sensível for sua carga de trabalho às terminações de instâncias, mais controle você poderá querer manter. Os exemplos a seguir descrevem cargas de trabalho com diferentes níveis de tolerância e as configurações recomendadas:

Exemplo 1: nós de banco de dados distribuídos (baixa tolerância)

Você executa um banco de dados distribuído em que cada instância do EC2 contém uma partição dos seus dados com um fator de replicação de 3. A perda de várias instâncias que contêm réplicas da mesma partição pode causar perda de dados ou tornar essa partição indisponível.

Desafio: o Auto Scaling pode encerrar instâncias mais rápido do que os dados podem ser replicados novamente em outros nós, e as terminações podem reduzir sua capacidade abaixo do necessário para manter seu fator de replicação.

Considere as seguintes configurações:

- Ative a proteção escalável de instâncias em todas as instâncias do banco de dados; remova-a programaticamente somente após confirmar que os dados foram replicados com segurança em outro lugar.

- Configure ganchos do ciclo de vida de encerramento com tempos limite estendidos em combinação com uma política de ciclo de vida da instância para permitir que a transferência de dados seja concluída.
- Defina uma política de manutenção de instâncias com uma porcentagem mínima de integridade de 100% para manter a capacidade necessária.

Exemplo 2: Processamento de trabalhos de longa duração (tolerância média)

Você tem uma fila do Amazon SQS que coleta mensagens recebidas para trabalhos de longa execução. Quando uma nova mensagem chega, uma instância do EC2 recupera a mensagem e inicia um trabalho que leva 3 horas para ser processado. Conforme a fila cresce, o Auto Scaling adiciona instâncias com base em suas políticas de escalabilidade. À medida que a fila diminui, o Auto Scaling encerra as instâncias.

Desafio: o Auto Scaling pode encerrar uma instância após 3 horas de processamento de uma tarefa, em vez de uma instância ociosa. O trabalho pode ser reiniciado em outra instância, mas você perde um progresso significativo.

Considere as seguintes configurações:

- Configure uma política de encerramento personalizada que priorize primeiro o encerramento de instâncias ociosas.
- Use ganchos de ciclo de vida de terminação para permitir que trabalhos em andamento sejam concluídos.
- Ative a proteção escalável de instâncias programaticamente quando uma instância iniciar um trabalho e remova-a quando o trabalho for concluído.

Exemplo 3: Frota de trabalhadores para ambientes de teste (alta tolerância)

Você executa uma frota de instâncias do EC2 que executam testes automatizados, trabalhos de CI/CD pipeline ou cargas de trabalho de desenvolvimento. Essas instâncias de trabalho extraem tarefas de uma fila e os resultados dos testes podem ser regenerados se um trabalho falhar.

Desafio: os trabalhos de teste podem ser interrompidos durante eventos de expansão, mas como os testes podem ser repetidos sem impacto, você deseja otimizar o custo e a simplicidade, em vez da disponibilidade sem interrupção.

Considere as seguintes configurações:

- Use ganchos de ciclo de vida de terminação para permitir que as solicitações em voo sejam concluídas.
- Considere usar instâncias spot com estratégia de alocação otimizada para capacidade para reduzir ainda mais os custos.

Suspender e retomar os processos do Amazon EC2 Auto Scaling

Este tópico explica como suspender e, depois, retomar um ou mais dos processos de escalabilidade para o grupo do Auto Scaling.

A suspensão de processos pode ser útil quando você precisa investigar ou solucionar um problema sem a interferência de políticas de escala ou ações programadas. Isso também ajuda a evitar que o Amazon EC2 Auto Scaling marque instâncias como não adequadas e as substitua enquanto você estiver fazendo alterações no seu grupo do Auto Scaling.

Tópicos

- Tipos de processos
- Considerações para suspender processos
- Suspender processos
- Retomar processos
- Entender como a suspensão de processos afeta outros processos

> **Note**
>
> Além de suspensões que você inicia, o Amazon EC2 Auto Scaling também pode suspender processos dos grupos do Auto Scaling que falharem repetidamente ao iniciar instâncias. Isso é conhecido como suspensão administrativa. Uma suspensão administrativa se aplica mais comumente a grupos do Auto Scaling que estão tentando iniciar instâncias por mais de 24 horas, mas não tiveram êxito. Você pode retomar os processos que foram suspensos pelo Amazon EC2 Auto Scaling por motivos administrativos.

Tipos de processos

O recurso suspender-retomar é compatível com os seguintes processos:

- `Launch`: adiciona instâncias ao grupo do Auto Scaling quando o grupo aumenta a escala horizontalmente ou quando o Amazon EC2 Auto Scaling opta por iniciar instâncias por outros motivos, p. ex., quando ele adiciona instâncias a um pool ativo.
- `Terminate`: remove instâncias do grupo do Auto Scaling quando o grupo reduz a escala horizontalmente ou quando o Amazon EC2 Auto Scaling opta por encerrar instâncias por outros motivos, p. ex., quando uma instância é encerrada por exceder a duração máxima de sua vida útil ou falhar em uma verificação de integridade.
- `AddToLoadBalancer`: adiciona instâncias ao grupo de destino do balanceador de carga anexado ou ao Classic Load Balancer quando elas são iniciadas. Para obter mais informações, consulte Use o Elastic Load Balancing para distribuir automaticamente o tráfego de entrada de aplicações no seu grupo do Auto Scaling..
- `AlarmNotification`— Aceita notificações de CloudWatch alarmes associados a políticas de escalabilidade dinâmica. Para obter mais informações, consulte Escalabilidade dinâmica para o Amazon EC2 Auto Scaling.
- `AZRebalance`: balanceia de maneira uniforme o número de instâncias do EC2 no grupo entre todas as zonas de disponibilidade especificadas quando o grupo fica desbalanceado, p. ex., quando uma zona de disponibilidade anteriormente indisponível volta para um estado íntegro. Para obter mais informações, consulte Atividades de rebalanceamento.
- `HealthCheck`: verifica a integridade das instâncias e marca uma instância como não íntegra se o Amazon EC2 ou o Elastic Load Balancing informar ao Amazon EC2 Auto Scaling que a instância não está íntegra. Esse processo pode substituir o status de integridade de uma instância que você definiu manualmente. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.
- `InstanceRefresh`: termina e substitui instâncias usando o recurso de atualização de instância. Para obter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling.
- `ReplaceUnhealthy` - encerra instâncias que estão marcadas como não íntegras e depois cria novas instâncias para substituí-las. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.
- `ScheduledActions`: executa as ações de escalabilidade programadas que você cria ou que são criadas para você na criação de um plano de escalabilidade da AWS Auto Scaling e ativação da escalabilidade preditiva. Para obter mais informações, consulte Escalabilidade programada para o Amazon EC2 Auto Scaling.

Considerações para suspender processos

Antes de suspender processos, considere o seguinte:

- A suspensão `AlarmNotification` permite que você interrompa temporariamente as políticas de rastreamento de metas, etapas e escalabilidade simples do grupo sem excluir as políticas de escalabilidade ou os alarmes associados. CloudWatch Para interromper temporariamente políticas individuais de escalabilidade, consulte Desabilitar uma política de escalabilidade para um grupo do Auto Scaling.
- Por exemplo, você pode suspender os processos `HealthCheck` e `ReplaceUnhealthy` para reinicializar instâncias sem que o Amazon EC2 Auto Scaling encerre as instâncias com base em suas verificações de integridade. Se você precisar que o Amazon EC2 Auto Scaling execute verificações de integridade nas instâncias restantes, use o recurso em espera em vez do recurso suspender-retomar. Para obter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.
- Se você suspender os processos `Launch` e `Terminate`, ou `AZRebalance`, e então fizer alterações em seu grupo do Auto Scaling, p. ex., desvinculando instâncias ou alterando as zonas de disponibilidade especificadas, seu grupo poderá ficar desbalanceado entre as zonas de disponibilidade. Se isso acontecer, depois que você retomar os processos suspensos, o Amazon EC2 Auto Scaling redistribuirá gradualmente as instâncias de modo uniforme entre as zonas de disponibilidade.
- Se você suspender o `Terminate` processo, ainda poderá forçar o encerramento das instâncias usando o delete-auto-scaling-groupcomando com a opção force delete.
- A suspensão do processo `Terminate` se aplica somente às instâncias que estão atualmente no estado `InService`. Ele não impede a terminação de instâncias em outros estados, como `Pending`, ou instâncias que não conseguem retomar adequadamente do modo de espera.
- O `RemoveFromLoadBalancerLowPriority` processo pode ser ignorado quando está presente em chamadas para descrever grupos de Auto Scaling usando o AWS CLI ou. SDKs Esse processo está defasado e é retido somente para fins de compatibilidade retroativa.

Suspender processos

Para suspender um processo de um grupo do Auto Scaling, use um dos métodos a seguir:

Console

Para suspender um processo

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Detalhes, escolha Configurações avançadas, Editar.
4. Em Suspended processes (Processos suspensos), escolha o processo a ser suspenso.
5. Selecione Atualizar.

AWS CLI

Use o seguinte comando suspend-processes para suspender processos individuais.

```
aws autoscaling suspend-processes --auto-scaling-group-name my-asg --scaling-
processes HealthCheck ReplaceUnhealthy
```

Para suspender todos os processos, omita a opção `--scaling-processes` como indicado abaixo.

```
aws autoscaling suspend-processes --auto-scaling-group-name my-asg
```

Retomar processos

Para retomar um processo suspenso para um grupo do Auto Scaling, use um dos métodos a seguir:

Console

Para retomar um processo suspenso

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.

3. Na guia Detalhes, escolha Configurações avançadas, Editar.
4. Em Suspended processes (Processos suspensos), escolha o processo suspenso.
5. Selecione Atualizar.

AWS CLI

Quando você estiver pronto para retomar um processo suspenso, use o seguinte comando resume-processes.

```
aws autoscaling resume-processes --auto-scaling-group-name my-asg --scaling-
processes HealthCheck
```

Para retomar todos os processos suspensos, omita a opção `--scaling-processes` como indicado abaixo.

```
aws autoscaling resume-processes --auto-scaling-group-name my-asg
```

Entender como a suspensão de processos afeta outros processos

As descrições a seguir explicam o que acontece quando tipos de processos individuais são suspensos.

Tópicos

- Launch está suspenso
- Terminate está suspenso
- AddToLoadBalancer está suspenso
- AlarmNotification está suspenso
- AZRebalance está suspenso
- HealthCheck está suspenso
- InstanceRefresh está suspenso
- ReplaceUnhealthy está suspenso
- ScheduledActions está suspenso
- Considerações adicionais

Launch está suspenso

- `AlarmNotification` ainda está ativo, mas seu grupo do Auto Scaling não pode iniciar atividades de aumento da escala na horizontal para alarmes que estejam violados.
- `ScheduledActions` está ativo, mas seu grupo do Auto Scaling não pode iniciar atividades de aumento da escala na horizontal para nenhuma ação de alarme que ocorra.
- `AZRebalance` deixa de rebalancear o grupo.
- `ReplaceUnhealthy` continua a encerrar instâncias não íntegras, mas não inicia instâncias substitutas. Quando você retomar o processo`Launch`, o Amazon EC2 Auto Scaling substituirá imediatamente todas as instâncias que ele encerrou enquanto `Launch` estava suspenso.
- `InstanceRefresh` não substitui as instâncias.

Terminate está suspenso

- `AlarmNotification` ainda está ativo, mas seu grupo do Auto Scaling não pode iniciar atividades de redução da escala na horizontal para alarmes que estejam violados.
- `ScheduledActions` está ativo, mas seu grupo do Auto Scaling não pode iniciar atividades de redução da escala na horizontal para nenhuma ação de alarme que ocorra.
- `AZRebalance` ainda fica ativo, mas não funciona corretamente. Ele pode iniciar novas instâncias sem encerrar as antigas. Isso pode fazer com que seu grupo do Auto Scaling cresça até 10% além de seu tamanho máximo, pois isso é permitido temporariamente durante atividades de rebalanceamento. Seu grupo do Auto Scaling poderá permanecer acima seu tamanho máximo até que você retome o processo `Terminate`.
- O `ReplaceUnhealthy` está inativo, mas não o `HealthCheck`. Quando o `Terminate` for reiniciado, o processo `ReplaceUnhealthy` começará a ser executado imediatamente. Se as instâncias foram marcadas como não íntegras enquanto o `Terminate` estava suspenso, elas serão substituídas imediatamente.
- `InstanceRefresh` não substitui as instâncias.

AddToLoadBalancer está suspenso

- O Amazon EC2 Auto Scaling executa as instâncias, mas não as adiciona ao grupo de destino do balanceador de carga ou ao Classic Load Balancer. Quando você retomar o processo `AddToLoadBalancer`, ele retomará a adição de instâncias ao balanceador de carga quando elas

forem iniciadas. No entanto, ele não adicionará as instâncias que foram iniciadas enquanto esse processo estava suspenso. É necessário registrar essas instâncias manualmente.

`AlarmNotification` está suspenso

- O Amazon EC2 Auto Scaling não invoca políticas de escalabilidade CloudWatch quando um limite de alarme é violado. Quando você retomar o `AlarmNotification`, o Amazon EC2 Auto Scaling levará em consideração as políticas com limites de alarme que estejam sendo violados no momento.

`AZRebalance` está suspenso

- Seu grupo do Amazon EC2 Auto Scaling não tenta redistribuir instâncias após determinados eventos. No entanto, se ocorrer um evento de expansão ou de redução, o processo de escalabilidade ainda tentará balancear as zonas de disponibilidade. Por exemplo, durante o aumento da escala na horizontal, ele executa instâncias na zona de disponibilidade com o menor número de instâncias. Se o grupo ficar desbalanceado enquanto `AZRebalance` estiver suspenso e você retomá-lo, o Amazon EC2 Auto Scaling tentará rebalancear o grupo. Ele chama primeiro o `Launch` e, depois, o `Terminate`.
- Os grupos de alta atividade não são afetados quando `AZRebalance` estiver suspenso.

`HealthCheck` está suspenso

- O Amazon EC2 Auto Scaling interrompe a marcação de instâncias com problemas de integridade como resultado das verificações de integridade do EC2 e do Elastic Load Balancing. Suas verificações personalizadas de integridade continuam funcionando corretamente. Depois que você suspender `HealthCheck`, se precisar, defina manualmente o estado de integridade das instâncias no seu grupo e faça com que o `ReplaceUnhealthy` as substitua.

`InstanceRefresh` está suspenso

- O Amazon EC2 Auto Scaling interrompe a substituição de instâncias como resultado de uma atualização de instância. Se houver uma atualização de instância em andamento, isso pausará a operação sem cancelá-la.

`ReplaceUnhealthy` está suspenso

- O Amazon EC2 Auto Scaling interrompe a substituição de instâncias que estão marcadas como não íntegras. As instâncias que falharem nas verificações de integridade do EC2 ou do Elastic Load Balancing ainda serão marcadas como não íntegras. Assim que você retomar o processo `ReplaceUnhealthy`, o Amazon EC2 Auto Scaling substituirá as instâncias que foram marcadas como não íntegras enquanto esse processo estava suspenso. O processo `ReplaceUnhealthy` chama `Terminate` primeiro e depois `Launch`.

`ScheduledActions` está suspenso

- O Amazon EC2 Auto Scaling não executa ações de escalabilidade que estejam programadas para execução durante o período de suspensão. Quando você retomar o `ScheduledActions`, o Amazon EC2 Auto Scaling considerará apenas ações programadas cuja programação ainda não tenha expirado.

Considerações adicionais

Além disso, quando `Launch` ou `Terminate` estiverem suspensos, os seguintes recursos podem não funcionar corretamente:

- Maximum instance lifetime (Tempo de vida máximo da instância): quando `Launch` ou `Terminate` estiverem suspensos, o recurso de tempo de vida máxima da instância não poderá substituir nenhuma instância.
- Interrupções de instâncias spot: se o `Terminate` estiver suspenso e seu grupo do Auto Scaling tiver instâncias spot, elas ainda poderão ser encerradas caso a capacidade spot não esteja mais disponível. Enquanto `Launch` estiver suspenso, o Amazon EC2 Auto Scaling não poderá iniciar instâncias substitutas de outro grupo de instâncias spot ou do mesmo grupo de instâncias spot quando ele estiver disponível novamente.
- Capacity Rebalancing (Rebalanceamento de capacidade): se `Terminate` estiver suspenso e você usar o rebalanceamento de capacidade para processar interrupções de instância spot, o serviço Amazon EC2 Spot ainda poderá encerrar instâncias caso a capacidade spot não esteja mais disponível. Se `Launch` estiver suspenso, o Amazon EC2 Auto Scaling não poderá iniciar instâncias substitutas de outro grupo de instâncias spot ou do mesmo grupo de instâncias spot quando ele estiver disponível novamente.

- Attaching and detaching instances (Anexação e desvinculação de instâncias): quando `Launch` e `Terminate` estiverem suspensos, você poderá desvincular instâncias anexadas ao seu grupo do Auto Scaling, mas enquanto `Launch` estiver suspenso, você não poderá anexar novas instâncias ao grupo.
- Standby instances (Instâncias em espera): quando `Launch` e `Terminate` estiverem suspensos, você poderá colocar uma instância no estado`Standby`, mas enquanto `Launch` estiver suspenso, você não poderá recolocar em serviço uma instância que esteja no estado `Standby`.

Monitore seus grupos do Amazon EC2 Auto Scaling

O monitoramento é uma parte importante da manutenção da confiabilidade, disponibilidade e desempenho do Amazon EC2 Auto Scaling e de suas soluções. Nuvem AWS AWS fornece as seguintes ferramentas de monitoramento para observar o Amazon EC2 Auto Scaling, relatar quando algo está errado e realizar ações automáticas quando apropriado:

Verificações de integridade

O Amazon EC2 Auto Scaling executa verificações de integridade nas instâncias do seu grupo do Auto Scaling. Se uma instância não passar na verificação de integridade, ela será marcada como não íntegra e será encerrada enquanto o Amazon EC2 Auto Scaling inicia uma nova instância para substituí-la. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

AWS Health Dashboard

Ele Health Dashboard exibe informações e também fornece notificações que são invocadas por alterações na integridade dos AWS recursos. As informações são apresentadas de duas formas: em um painel que mostra eventos recentes e futuros organizados por categoria e em um log de eventos completo que mostra todos os eventos dos últimos 90 dias. Para obter mais informações, consulte Health Dashboard notificações para Amazon EC2 Auto Scaling.

CloudTrail

Com AWS CloudTrail, você pode rastrear as chamadas feitas para a API do Amazon EC2 Auto Scaling por ou em nome de você. Conta da AWS CloudTrail armazena as informações em arquivos de log no bucket do Amazon S3 que você especificar. Você pode usar esses arquivos de log para monitorar a atividade de seus grupos do Auto Scaling. Os logs incluem quais solicitações foram feitas, os endereços IP de onde as solicitações vieram, quem fez a solicitação, quando a solicitação foi feita e assim por diante. Para obter mais informações, consulte Registre chamadas da API Amazon EC2 Auto Scaling com AWS CloudTrail.

> **Coleta de logs nas instâncias do Amazon EC2**
>
> Você pode usar CloudWatch para coletar registros dos sistemas operacionais para suas instâncias do EC2. Para obter mais informações, consulte Coletar métricas e registros de instâncias do Amazon EC2 e servidores locais com o CloudWatch agente e Visualizar dados de log enviados para CloudWatch Logs no Guia do usuário da Amazon CloudWatch .

Para obter informações sobre outros AWS serviços que podem ajudá-lo a registrar e coletar dados sobre suas cargas de trabalho, consulte o guia de registro e monitoramento para proprietários de aplicativos na Orientação AWS prescritiva.

Amazon CloudWatch

CloudWatch A Amazon ajuda você a analisar registros e, em tempo real, monitorar as métricas de seus AWS recursos e aplicativos hospedados. Você pode coletar e rastrear métricas, criar painéis personalizados e definir alarmes que o notificam ou que realizam ações quando uma métrica especificada atinge um limite definido. Por exemplo, é possível ser notificado quando a atividade da rede é repentinamente maior ou menor do que o valor esperado de uma métrica. Para obter mais informações sobre o uso do serviço de monitorar as métricas dos grupos do Auto Scaling e instâncias, consulte Monitore CloudWatch métricas para seus grupos e instâncias do Auto Scaling.

CloudWatch também monitora as métricas de uso AWS da API para o Amazon EC2 Auto Scaling. Você pode usar essas métricas para configurar alarmes que alertem quando o volume de chamadas da API violar um limite definido por você. Para obter mais informações, consulte as métricas de AWS uso no Guia CloudWatch do usuário da Amazon.

AWS Compute Optimizer

O Compute Optimizer informa as recomendações de instâncias do Amazon EC2 que podem ajudar a decidir se deseja passar para um novo tipo de instância. Ele analisa se o tipo de instância do grupo do Auto Scaling está em condições ideais e gera recomendações de otimização para reduzir o custo e melhorar o desempenho de suas workloads.. Para obter mais informações, consulte Obtenha recomendações de tipo de instância com AWS Compute Optimizer.

Amazon EventBridge

EventBridge A Amazon é um serviço de ônibus de eventos sem servidor que facilita a conexão de seus aplicativos com dados de várias fontes. EventBridge fornece um fluxo de dados em tempo real de seus próprios aplicativos, aplicativos Software-as-a-Service (SaaS) e AWS serviços e encaminha esses dados para destinos como o Lambda. Isso permite monitorar eventos que ocorrem em serviços e criar arquiteturas orientadas a eventos. Para obter mais informações, consulte Use EventBridge para lidar com eventos do Auto Scaling.

AWS Security Hub CSPM

Use AWS Security Hub CSPM para monitorar o uso do Amazon EC2 Auto Scaling já que ele se refere às melhores práticas de segurança. O Security Hub CSPM usa controles de segurança de detetive para avaliar as configurações de recursos e os padrões de segurança para ajudá-lo a cumprir várias estruturas de conformidade. Para obter mais informações sobre o uso do Security Hub CSPM para avaliar os recursos do Amazon EC2 Auto Scaling, consulte Controles do Amazon EC2 Auto Scaling no Guia do usuário.AWS Security Hub

Amazon Simple Notification Service

Você pode configurar grupos do Auto Scaling para enviar notificações do Amazon SNS sempre que o Amazon EC2 Auto Scaling iniciar ou terminar instâncias. Para obter mais informações, consulte Configurar notificações do Amazon SNS para o Amazon EC2 Auto Scaling.

Verificações de integridade para instâncias em um grupo do Auto Scaling

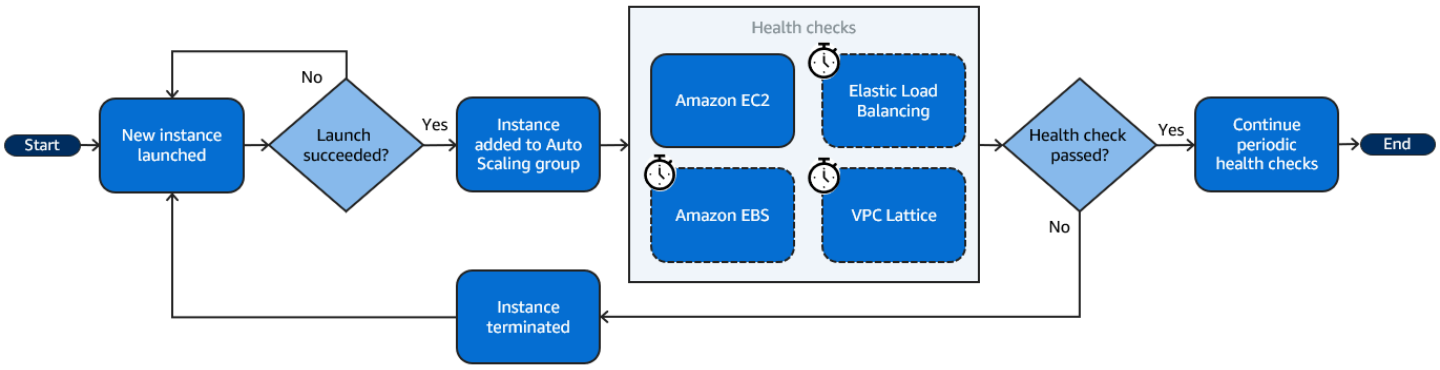
O Amazon EC2 Auto Scaling monitora continuamente o status de integridade das instâncias em um grupo do Auto Scaling para manter a capacidade desejada.

Todas as instâncias do seu grupo do Auto Scaling começam com um status`Healthy`. Supõe-se que as instâncias estejam íntegras, a menos que o Amazon EC2 Auto Scaling receba uma notificação informando do contrário. Ele pode receber notificações de várias fontes quando uma instância se torna inadequada e precisa ser substituída. Essas fontes incluem o seguinte:

- Amazon EC2
- Elastic Load Balancing
- Estrutura de PVC
- Amazon EBS
- Verificações de integridade personalizadas que você define

Quando o Amazon EC2 Auto Scaling determina que uma `InService` instância não está íntegra, ele a substitui por uma nova instância. A nova instância é iniciada usando as configurações atuais do grupo do Auto Scaling e seu modelo de execução associado ou configuração de execução.

O fluxograma a seguir ilustra o processo de execução de uma nova instância em um grupo do Auto Scaling. Ele começa com a execução da instância. Se a execução for bem-sucedida, a instância será adicionada ao grupo do Auto Scaling. Em seguida, o Amazon EC2 Auto Scaling realiza verificações de integridade na instância usando as verificações de estado integradas do Amazon EC2 e, após um período de carência, quaisquer verificações de integridade opcionais que você tenha ativado para o grupo. Essas verificações de integridade continuam periodicamente. Se alguma das verificações de integridade falhar, a instância será substituída.



Instâncias não íntegras também podem ocorrer quando uma instância é terminada inesperadamente, por exemplo, devido a uma interrupção da instância spot ou terminação manual por um usuário. Novamente, o Amazon EC2 Auto Scaling iniciará automaticamente uma instância de substituição nesses casos para manter a capacidade desejada.

Conteúdo

- Sobre as verificações de integridade do seu grupo do Auto Scaling
- Definir um período de carência da verificação de integridade para um grupo do Auto Scaling
- Monitore as instâncias do Auto Scaling com volumes do Amazon EBS com problemas usando verificações de integridade
- Configuração de uma verificação de integridade personalizada para seu grupo do Auto Scaling
- Veja o motivo das falhas na verificação de integridade
- Solução de problemas de instâncias não íntegras no Amazon EC2 Auto Scaling

Sobre as verificações de integridade do seu grupo do Auto Scaling

Este tópico fornece uma visão geral dos tipos de verificação de integridade disponíveis e descreve as principais considerações para integrar as verificações de integridade do Amazon EC2 Auto Scaling às suas aplicações.

Conteúdo

- Tipo de verificação de integridade
- Verificações de integridade do Amazon EC2
- Verificações de saúde do Elastic Load Balancing
- Verificações de integridade do VPC Lattice
- Como o Amazon EC2 Auto Scaling minimiza o tempo de inatividade
- Verificações de integridade para instâncias em uma associação de alta atividade
- Considerações sobre a verificação de integridade

Tipo de verificação de integridade

O Amazon EC2 Auto Scaling pode determinar o status de integridade de uma instância da `InService` usando uma ou mais das seguintes verificações de integridade:

Tipo de verificação de integridade	O que ele verifica
Verificações de status do Amazon EC2 e eventos programados	• Verifica se a instância está em execução • Verifica se há problemas subjacentes de hardware ou software capazes de prejudicar a instância Esse é o tipo padrão de verificação de integridade para um grupo do Auto Scaling.
Verificações de saúde do Elastic Load Balancing	• Verifica se o balanceador de carga relata a instância como íntegra, confirmando se a instância está disponível para processar solicitações. Para executar esse tipo de verificação de integridade, é necessário habilitá-lo para seu grupo do Auto Scaling.
Verificações de integridade do VPC Lattice	• Verifica se o VPC Lattice relata a instância como íntegra, confirmando se a instância está disponível para lidar com solicitações

Tipo de verificação de integridade	O que ele verifica
	Para executar esse tipo de verificação de integridade, é necessário habilitá-lo para seu grupo do Auto Scaling.
Verificações de saúde do Amazon EBS	• Verifica se os volumes do EBS estão acessíveis e passa I/O nas verificações de status. Para executar esse tipo de verificação de integridade, é necessário habilitá-lo para seu grupo do Auto Scaling.
Verificações de integridade personalizadas	• Verifica se há outros problemas que possam indicar problemas de integridade da instância de acordo com suas verificações de integridade personalizadas

Verificações de integridade do Amazon EC2

Depois que uma instância é executada, ela é anexada ao grupo do Auto Scaling e entra no estado`InService`. Para obter mais informações sobre os diferentes status do ciclo de vida de instâncias em um grupo do Auto Scaling, consulte Ciclo de vida das instâncias do Amazon EC2 Auto Scaling.

O Amazon EC2 Auto Scaling verifica periodicamente o status de integridade de todas as instâncias no grupo do Auto Scaling para garantir que elas estejam em execução e em boas condições.

Verificações do status

O Amazon EC2 Auto Scaling usa os resultados das verificações de status da instância do Amazon EC2 para determinar o status de integridade de uma instância. Se a instância estiver em qualquer estado do Amazon EC2 diferente de `running` ou se o status para as verificações de status mudar para `impaired`, o Amazon EC2 Auto Scaling considerará a instância como não íntegra e a substituirá. Isso inclui quando a instância tenha qualquer um dos seguintes estados:

- `stopping`
- `stopped`
- `shutting-down`
- `terminated`

As verificações de status do Amazon EC2 não exigem nenhuma configuração especial e estão sempre habilitadas. Para obter mais informações, consulte Tipos de verificações de status no Manual do usuário do Amazon EC2 para instâncias do Linux.

> **Important**
>
> O Amazon EC2 Auto Scaling permite que essas verificações de status falhem ocasionalmente sem realizar qualquer ação. Quando uma verificação de status falha, o Amazon EC2 Auto Scaling espera alguns AWS minutos para corrigir o problema. Ele não marca imediatamente uma instância `Unhealthy` quando seu status para as verificações de status se torna `impaired`. Além disso, o EC2 Auto Scaling não marcará a instância como `Unhealthy` se uma verificação de status retornar `insufficient-data`.
> No entanto, se o Amazon EC2 Auto Scaling detectar que uma instância não está mais no estado`running`, essa situação será tratada como uma falha imediata. Neste caso, marca imediatamente a instância `Unhealthy` e a substitui.

Eventos agendados

O Amazon EC2 pode agendar ocasionalmente eventos em suas instâncias que serão executados após um carimbo de data/hora específico. Para obter mais informações, consulte Eventos programados para sua instância no Guia do usuário do Amazon EC2 para instâncias do Linux.

Se uma de suas instâncias for afetada por um evento programado, o Amazon EC2 Auto Scaling considerará a instância como não íntegra e a substituirá. A instância não começa a ser encerrada até que a data e a hora especificadas no carimbo de data/hora sejam atingidas.

Verificações de saúde do Elastic Load Balancing

Quando você habilita as verificações de integridade do Elastic Load Balancing para seu grupo do Auto Scaling, o Amazon EC2 Auto Scaling pode usar os resultados dessas verificações de integridade para determinar o status de integridade de uma instância.

Antes de ativar as verificações de integridade do Elastic Load Balancing para o seu grupo do Auto Scaling, é necessário configurar um balanceador de carga do Elastic Load Balancing e configurar uma verificação de integridade para ele a fim de determinar se as instâncias estão íntegras. Para obter mais informações, consulte Para anexar um balanceador de carga do Elastic Load Balancing a uma camada.

Depois de anexar o balanceador de carga ao seu grupo do Auto Scaling, ocorre o seguinte:

- O Amazon EC2 Auto Scaling registrará as instâncias no grupo do Auto Scaling com o balanceador de carga.
- Depois que uma instância termina de registrar, ela entra no estado `InService` e fica disponível para uso com o balanceador de carga.

Por padrão, o Amazon EC2 Auto Scaling ignora os resultados das verificações de integridade do Elastic Load Balancing. Depois de fazer isso, quando o Elastic Load Balancing relata uma instância registrada como`Unhealthy`, o Amazon EC2 Auto Scaling marca a instância `Unhealthy` na próxima verificação de integridade periódica e a substitui.

Se a drenagem da conexão (atraso de cancelamento de registro) estiver habilitada para seu balanceador de carga, o Amazon EC2 Auto Scaling aguardará que as solicitações em andamento sejam concluídas ou que o tempo limite máximo expire antes de terminar instâncias não íntegras.

> **Note**
>
> Para obter instruções sobre como anexar o balanceador de carga e ativar as verificações de integridade do Elastic Load Balancing para seu grupo do Auto Scaling, consulte Este tópico descreve como anexar um balanceador de carga Elastic Load Balancing ao grupo do Auto Scaling..
> Quando você habilita as verificações de integridade do Elastic Load Balancing para um grupo, o Amazon EC2 Auto Scaling pode substituir instâncias que o Elastic Load Balancing relata como não íntegras, mas somente depois que o balanceador de carga estiver no `InService` estado. Para obter mais informações, consulte Verificar o status do anexo de seu balanceador de carga.

Verificações de integridade do VPC Lattice

Por padrão, o Amazon EC2 Auto Scaling ignora os resultados das verificações de integridade da VPC Lattice. Opcionalmente, você pode ativar essas verificações de integridade para seu grupo do Auto Scaling. Depois de fazer isso, quando o VPC Lattice relata uma instância registrada como`Unhealthy`, o Amazon EC2 Auto Scaling marca a instância `Unhealthy` na próxima verificação de integridade periódica e a substitui. O processo de registrar instâncias e, em seguida, verificar sua integridade funciona da mesma forma que as verificações de integridade do Elastic Load Balancing.

> **Note**
>
> Para obter instruções sobre como anexar o grupo-alvo do VPC Lattice e ativar as verificações de integridade do VPC Lattice para o grupo do Auto Scaling, consulte Anexe um grupo de destino VPC Lattice ao seu grupo do Auto Scaling.
> Quando você habilita verificações de integridade do VPC Lattice para um grupo, o Amazon EC2 Auto Scaling pode substituir instâncias que o VPC Lattice relata como não íntegras, mas somente depois que o grupo de destino estiver no estado `InService`. Para obter mais informações, consulte Verifique o status do anexo do grupo de destino do VPC Lattice.

Como o Amazon EC2 Auto Scaling minimiza o tempo de inatividade

Por padrão, as novas instâncias são provisionadas ao mesmo tempo em que as instâncias existentes são terminadas, o que pode impedir que novas solicitações sejam aceitas até que as novas instâncias estejam totalmente operacionais.

Se o Amazon EC2 Auto Scaling determinar que alguma instância não está mais em execução (ou `Unhealthy` foi marcada set-instance-healthcom o comando), ele a substituirá imediatamente. No entanto, se houver outras instâncias não íntegras, o Amazon EC2 Auto Scaling usará a abordagem a seguir para se recuperar de falhas. Esta abordagem minimiza qualquer tempo de inatividade que possa ocorrer devido a problemas temporários ou verificações de integridade mal configuradas.

- Se houver uma ação de escalabilidade em andamento e seu grupo do Auto Scaling estiver abaixo da capacidade desejada em 10% ou mais, o Amazon EC2 Auto Scaling aguarda a atividade de escalabilidade em andamento antes de substituir as instâncias não íntegras.
- Ao aumentar a escala horizontalmente, o Amazon EC2 Auto Scaling aguarda que as instâncias passem por uma verificação de integridade inicial. Ele também aguarda que o aquecimento padrão de instância seja concluído para garantir que as novas instâncias estejam prontas.
- Depois que as instâncias terminarem de aquecer e o grupo tiver aumentado para acima de 90% da capacidade desejada, o Amazon EC2 Auto Scaling substituirá as instâncias não íntegras da seguinte maneira:
  - O Amazon EC2 Auto Scaling substitui apenas até 10% da capacidade desejada do grupo por vez. Ele faz isso até que todas as instâncias não íntegras sejam substituídas.
  - Ao substituir instâncias, ele espera que as novas instâncias passem por uma verificação de integridade inicial. Ele também aguarda que o aquecimento padrão de instância seja concluído antes de continuar.

Note

- Se o tamanho de um grupo do Auto Scaling for suficientemente pequeno para que o valor resultante de 10% seja menor que um, o Amazon EC2 Auto Scaling substituirá cada uma das instâncias não íntegras de cada vez. Isso pode resultar em tempo de inatividade para o grupo.
- É possível modificar o valor padrão de 10% definindo uma política de manutenção de instâncias para alterar a taxa na qual o Auto Scaling substitui instâncias não íntegras. No entanto, o Auto Scaling ainda pode limitar a taxa de marcação de instâncias como não íntegras.

  Por exemplo, se todas as instâncias em um grupo do Auto Scaling forem relatadas como não íntegras pelas verificações de integridade do Elastic Load Balancing e o balanceador de carga estiver no estado `InService`, o Amazon EC2 Auto Scaling poderá marcar menos instâncias não íntegras de cada vez. Isso pode resultar em muito menos instâncias substituídas por vez do que os 10% aplicados em outros cenários. Isso fornece tempo para resolver o problema sem que o Amazon EC2 Auto Scaling encerre automaticamente todo o grupo.

Verificações de integridade para instâncias em uma associação de alta atividade

Saiba mais sobre as verificações de integridade do Amazon EC2 Auto Scaling para instâncias em um pool ativo. Para obter mais informações, consulte Visualizar o status e o motivo de falhas da verificação de integridade.

Considerações sobre a verificação de integridade

Veja a seguir algumas considerações ao usar as verificações de integridade do Amazon EC2 Auto Scaling.

- Se precisar que algo aconteça na instância que está sendo terminada ou na instância que está iniciando, você poderá usar ganchos do ciclo de vida. Esses hooks permitem que você execute uma ação personalizada à medida que o Amazon EC2 Auto Scaling inicia ou encerra instâncias. Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.
- O Amazon EC2 Auto Scaling não fornece um modo de remover as verificações de status e eventos programados do Amazon EC2 das verificações de integridade. Se você não quiser que

as instâncias sejam substituídas, recomendamos suspender o processo `ReplaceUnhealthy` e `HealthCheck` para grupos do Auto Scaling individuais. Para obter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.

- Para redefinir manualmente o status de integridade de uma instância não `Healthy` íntegra, tente usar o set-instance-healthcomando. Se você receber um erro, provavelmente a instância já está encerrando. Geralmente, redefinir o status de integridade de uma instância `Healthy` com o set-instance-healthcomando só é útil nos casos em que o `ReplaceUnhealthy` processo ou o `Terminate` processo estão suspensos.
- Se você precisar solucionar problemas de uma instância sem a interferência de verificações de integridade, poderá colocar a instância no estado `Standby`. O Amazon EC2 Auto Scaling não executa verificações de integridade em instâncias que estão no estado `Standby`. Para obter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.
- Quando a instância é encerrada, qualquer endereço IP elástico é dissociado e não é automaticamente associado à nova instância. É necessário associar manualmente os endereços IP elásticos à nova instância ou fazer isso automaticamente com uma solução baseada em gancho do ciclo de vida. Para obter mais informações, consulte Endereços IP elásticos no Guia do usuário do Amazon EC2.
- Da mesma forma, quando sua instância é terminada, seus volumes de EBS anexados são desvinculados (ou excluídos, dependendo do atributo `DeleteOnTermination`). É necessário anexar manualmente esses volumes do EBS à nova instância ou fazer isso automaticamente com uma solução baseada em gancho do ciclo de vida. Para obter mais informações, consulte Associar um volume do Amazon EBS a uma instância no Guia do usuário do Amazon EBS.

Definir um período de carência da verificação de integridade para um grupo do Auto Scaling

Quando uma verificação de integridade do Amazon EC2 Auto Scaling determina que uma `InService` instância não está íntegra, ela a substitui por uma nova instância. O período de carência da verificação de integridade especifica o tempo mínimo (em segundos) que uma nova instância será mantida em serviço antes de ser terminada, caso não esteja íntegra.

Um exemplo de caso de uso pode ser a exigência de que o Amazon EC2 Auto Scaling evite realizar ações se as verificações de integridade do Elastic Load Balancing falharem porque uma instância ainda está sendo inicializada. As verificações de integridade do Elastic Load Balancing são executadas em paralelo, começando quando a instância é registrada no balanceador de carga. O período de carência evita que o Amazon EC2 Auto Scaling marque suas instâncias recém-iniciadas

`Unhealthy` e as encerre desnecessariamente se elas não passarem imediatamente nessas verificações de integridade após entrarem no `InService` estado.

Por padrão, o período de carência da verificação de integridade é de 300 segundos quando você cria um grupo do Auto Scaling. Seu valor padrão é 0 segundos quando você cria um grupo de Auto Scaling usando o AWS CLI ou um SDK. O valor 0 desativa um período de carência da verificação de integridade.

Definir um valor muito alto reduz a eficácia das verificações de integridade do Amazon EC2 Auto Scaling. Se você usar ganchos do ciclo de vida para iniciar a instância, poderá definir o valor do período de carência da verificação de integridade como 0. Com ganchos do ciclo de vida, o Amazon EC2 Auto Scaling fornece uma maneira de garantir que as instâncias sejam sempre inicializadas antes de entrarem no estado `InService`. Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

O período de carência se aplica às seguintes instâncias:

- Instâncias recém-lançadas
- Instâncias que são colocadas de volta em serviço após estarem em espera
- Instâncias que você anexa manualmente ao grupo

> **Important**
>
> Durante o período de carência da verificação de integridade, se o Amazon EC2 Auto Scaling detectar que uma instância não está mais no estado Amazon EC2, `running` ele marcará imediatamente a instância `Unhealthy` e a substituirá. Por exemplo, se você interromper uma instância em um grupo do Auto Scaling, ela será marcada `Unhealthy` e substituída.

Definir um período de carência da verificação de integridade para um grupo

Definir um período de carência da verificação de integridade para grupos do Auto Scaling existentes.

Console

Para modificar o período de carência da verificação de integridade para um grupo novo ()

Quando você cria o grupo do Auto Scaling, na página Configure advanced options (Configurar opções avançadas), em Health checks (Verificações de integridade), Health check grace period

(Período de carência da verificação de integridade), insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`.

AWS CLI

Para modificar o período de carência da verificação de integridade para um grupo novo ()

Adicione a `--health-check-grace-period` opção ao create-auto-scaling-groupcomando. O exemplo a seguir configura o período de carência da verificação de integridade com um valor de *60* segundos para um novo grupo do Auto Scaling denominado *my-asg*.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-grace-period 60 ...
```

Console

Para modificar o período de carência da verificação de integridade para um grupo existente ()

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a mesma Região da AWS na qual você criou o grupo do Auto Scaling.
3. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
4. Na guia Detalhes, escolha Verificações de integridade, Editar.
5. Em Health check grace period (Período de carência da verificação de integridade), insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`.
6. Selecione Atualizar.

AWS CLI

Para modificar o período de carência da verificação de integridade para um grupo existente ()

Adicione a `--health-check-grace-period` opção ao update-auto-scaling-groupcomando. O exemplo a seguir configura o período de carência da verificação de integridade com um valor de *120* segundos para um grupo do Auto Scaling existente denominado *my-asg*.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-grace-period 120
```

> **Note**
>
> Também é altamente recomendável definir o tempo de aquecimento padrão da instância para o grupo do Auto Scaling. Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Monitore as instâncias do Auto Scaling com volumes do Amazon EBS com problemas usando verificações de integridade

Você pode ativar as verificações de integridade do Amazon EBS para seu grupo do Auto Scaling para garantir que o Amazon EC2 Auto Scaling monitore todo o sistema no qual sua aplicação é executada.

Depois que você ativar essas verificações de integridade, o Amazon EC2 Auto Scaling receberá os resultados das verificações de estado do Amazon EC2 realizadas nos volumes EBS anexados de uma instância. Se um volume não estiver acessível ou não passar nas verificações de I/O status, a verificação de integridade falhará e a instância correspondente será considerada não íntegra. Quando o Amazon EC2 Auto Scaling determina que uma instância não está íntegra, ele a substitui por uma nova instância.

Este tópico pressupõe que você esteja familiarizado com as verificações de estado do EBS anexadas. Caso contrário, consulte a seção Verificações de estado do EBS anexado do Guia do usuário do Amazon EC2 para obter mais informações. O tópico a seguir descreve como ativar as verificações de integridade do Amazon EC2 Auto Scaling que dependem das verificações de estado do EBS anexado.

> **Note**
>
> Você pode ativar as verificações de integridade do Amazon EBS para todos os seus grupos do Auto Scaling. No entanto, essas verificações de integridade só estão disponíveis para instâncias criadas no Nitro System da AWS.

Ativar as verificações de integridade do Amazon EBS para um grupo

Você pode ativar as verificações de integridade do Amazon EBS para grupos do Auto Scaling novos e existentes.

Console

Ativação das verificações de integridade do Amazon EBS para um novo grupo

Ao criar o grupo do Auto Scaling, na página Configurar opções avançadas, para Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Amazon EBS. Em seguida, para Período de carência da verificação de integridade, insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

AWS CLI

Ativação das verificações de integridade do Amazon EBS para um novo grupo

Adicione a `--health-check-type` opção ao create-auto-scaling-groupcomando. O exemplo a seguir especifica **`EBS`** para a opção `--health-check-type` para um novo grupo do Auto Scaling chamado *`my-asg`*.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "EBS" --health-check-grace-period 60 ...
```

Você pode especificar vários valores para o elemento `--health-check-type`. Por exemplo, para adicionar os tipos de verificações de integridade do Amazon EBS e do Elastic Load Balancing, use o seguinte comando.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
```

```
  --health-check-type "EBS,ELB" --health-check-grace-period 60 ...
```

Os nomes dos valores diferenciam maiúsculas de minúsculas.

Console

Ativação das verificações de integridade do Amazon EBS para um grupo existente

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a mesma Região da AWS na qual você criou o grupo do Auto Scaling.
3. Marque a caixa de seleção ao lado de um grupo existente.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
4. Na guia Detalhes, escolha Verificações de integridade, Editar.
5. Em Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Amazon EBS.
6. Em Período de carência da verificação de integridade, insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.
7. Selecione Atualizar.

AWS CLI

Ativação das verificações de integridade do Amazon EBS para um grupo existente

Adicione a `--health-check-type` opção ao update-auto-scaling-groupcomando. O exemplo a seguir especifica **`EBS`** para a opção `--health-check-type` para um grupo do Auto Scaling existente chamado *`my-asg`*.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "EBS" --health-check-grace-period 60
```

Para usar vários tipos de verificações de integridade, você pode especificar vários valores (por exemplo, **`EBS,ELB`**) para a opção `--health-check-type`.

Os nomes dos valores diferenciam maiúsculas de minúsculas.

Desativar as verificações de integridade do Amazon EBS para um grupo do Auto Scaling

O tópico a seguir descreve como desativar as verificações de integridade do Amazon EBS para um grupo do Auto Scaling. Se você não precisar mais das verificações de integridade do Amazon EBS, use o procedimento a seguir para desativá-las.

Console

Desativação das verificações de integridade do Amazon EBS para um novo grupo

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado de um grupo existente.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Detalhes, escolha Verificações de integridade, Editar.
4. Em Verificações de integridade, Tipos adicionais de verificação de integridade, desmarque Ativar verificações de integridade do Amazon EBS.
5. Selecione Atualizar.

AWS CLI

Desativação das verificações de integridade do Amazon EBS para um novo grupo

Para atualizar as verificações de saúde em um grupo do Auto Scaling para que ele não use mais as verificações de saúde do Amazon EBS, use o comando. update-auto-scaling-group Inclua a opção `--health-check-type` e um valor de **`EC2`**.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "EC2"
```

Para desativar as verificações de integridade do Amazon EBS sem desativar outros tipos de verificação de integridade (como o Elastic Load Balancing), é necessário especificá-las em vez de **`EC2`**. Por exemplo, para verificações de integridade do Elastic Load Balancing, especifique **`ELB`** para a opção `--health-check-type`.

Os nomes dos valores diferenciam maiúsculas de minúsculas.

Configuração de uma verificação de integridade personalizada para seu grupo do Auto Scaling

Você pode usar verificações de integridade personalizadas para complementar as opções de verificação de integridade existentes fornecidas pelo Amazon EC2 Auto Scaling. Ao combinar verificações de integridade personalizadas com outros tipos de verificação de integridade, é possível criar um sistema abrangente de monitoramento de integridade adaptado às necessidades da sua aplicação.

Para começar, crie testes personalizados para verificar se as instâncias em seu grupo do Auto Scaling estão funcionando corretamente e podem lidar com o tráfego de entrada. Se a verificação de integridade que você configurar detectar que uma instância não está respondendo, marque essa instância específica como `Unhealthy`, o que faz com que o Amazon EC2 Auto Scaling a substitua imediatamente.

Você pode enviar as informações de integridade da instância diretamente ao Amazon EC2 Auto Scaling usando a AWS CLI ou um SDK. Os exemplos a seguir mostram como usar o AWS CLI para configurar o status de integridade de uma instância e depois verificar o status de integridade da instância.

Use o set-instance-healthcomando a seguir para definir o status de integridade da instância especificada como**`Unhealthy`**.

```
aws autoscaling set-instance-health --instance-id i-1234567890abcdef0 --health-
status Unhealthy
```

Por padrão, esse comando respeita o período de carência da verificação de integridade. Porém, é possível substituir esse comportamento e não respeitar o período de carência incluindo a opção`--no-should-respect-grace-period`.

Use o describe-auto-scaling-groupscomando a seguir para verificar se o status de integridade da instância é`Unhealthy`.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-names my-asg
```

O exemplo a seguir é de uma resposta que mostra que o status de integridade da instância é `Unhealthy` e que a instância está sendo encerrada.

```
{
    "AutoScalingGroups": [
        {
            ....
            "Instances": [
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-1234567890abcdef0"
                    },
                    "InstanceId": "i-1234567890abcdef0",
                    "InstanceType": "t2.micro",
                    "HealthStatus": "Unhealthy",
                    "LifecycleState": "Terminating"
                },
                ...
            ]
        }
    ]
}
```

Veja o motivo das falhas na verificação de integridade

Usando o procedimento a seguir, você pode exibir informações sobre todas as instâncias substituídas devido a uma verificação de integridade.

Por padrão, o Amazon EC2 Auto Scaling cria uma nova atividade de escalabilidade para encerrar a instância não íntegra e, em seguida, finaliza-a. Enquanto a instância estiver sendo encerrada, outra atividade de escalonamento iniciará uma nova instância. Você pode alterar esse comportamento para iniciar uma nova instância primeiro usando uma política de manutenção de instâncias. Para obter mais informações, consulte Políticas de manutenção de instância.

Console

Para visualizar o motivo das falhas de verificação de integridade ()

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status mostra se o seu grupo do Auto Scaling iniciou ou terminou instâncias com êxito.

   Se ele terminou quaisquer instâncias não íntegras, a coluna Cause (Causa) mostrará a data e a hora do término e o motivo da falha na verificação de integridade. Por exemplo, `.At 2022-05-14T20:11:53Z an instance was taken out of service in response to a user health-check` Essa mensagem indica que uma verificação de integridade personalizada marcou a instância como não íntegra.

   Para obter ajuda com falhas na verificação de integridade, consulte Solução de problemas de instâncias não íntegras no Amazon EC2 Auto Scaling.

AWS CLI

Para visualizar o motivo das falhas de verificação de integridade ()

Use o seguinte comando describe-scaling-activities:

```
aws autoscaling describe-scaling-activities --auto-scaling-group-name my-asg
```

A seguir, um exemplo de resposta, no que `Cause` contém o motivo da falha na verificação de integridade.

```
{
  "Activities": [
    {
      "ActivityId": "4c65e23d-a35a-4e7d-b6e4-2eaa8753dc12",
      "AutoScalingGroupName": "my-asg",
      "Description": "Terminating EC2 instance: i-04925c838b6438f14",
```

```
      "Cause": "At 2021-04-01T21:48:35Z an instance was taken out of service in
 response to a user health-check.",
      "StartTime": "2021-04-01T21:48:35.859Z",
      "EndTime": "2021-04-01T21:49:18Z",
      "StatusCode": "Successful",
      "Progress": 100,
      "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-
west-2a\"...}",
      "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
    },
...
  ]
}
```

Para obter uma descrição dos campos na saída, consulte Atividade na Referência da API do Amazon EC2 Auto Scaling.

Para descrever as atividades de escalabilidade após a exclusão do grupo Auto Scaling, adicione `--include-deleted-groups` a opção ao describe-scaling-activitiescomando.

Solução de problemas de instâncias não íntegras no Amazon EC2 Auto Scaling

A seguir estão mensagens de erro apresentadas pelo Amazon EC2 Auto Scaling, as possíveis causas e as etapas que você pode tomar para resolver o problema.

Para recuperar uma mensagem de erro, consulte Veja o motivo das falhas na verificação de integridade.

Mensagens de erro

- Uma instância foi retirada de serviço em resposta a uma falha de verificação de status de instância do EC2
- Uma instância foi retirada de serviço em resposta a uma verificação de integridade do EC2 que indicou que ela tinha sido terminada ou interrompida
- Uma instância foi retirada de serviço em resposta a uma falha na verificação de integridade do sistema ELB
- Recursos adicionais do

Uma instância foi retirada de serviço em resposta a uma falha de verificação de status de instância do EC2

Problema: instâncias do Auto Scaling falham nas verificações de status do Amazon EC2.

Causa 1: se houver problemas que fazem com que o Amazon EC2 considere as instâncias do grupo do Auto Scaling prejudicadas, o Amazon EC2 Auto Scaling substituirá automaticamente as instâncias prejudicadas como parte da verificação de integridade.

Solução 1:: Quando uma verificação de status de instância falha, geralmente você precisa lidar com o problema por conta própria fazendo alterações de configuração da instância até que a aplicação não apresente mais problemas. Para resolver esse problema, siga estas etapas:

1. Crie manualmente uma instância do Amazon EC2 que não faça parte do grupo do Auto Scaling e investigue o problema. Para obter ajuda geral na investigação de instâncias comprometidas, consulte Solução de problemas em instâncias com falha nas verificações de status no Guia do usuário do Amazon EC2.
2. Depois de confirmar que sua instância foi executada com êxito e está íntegra, implante uma nova configuração de instância, livre de erros, no grupo do Auto Scaling.
3. Exclua a instância criada para evitar cobranças contínuas na conta da AWS .

Uma instância foi retirada de serviço em resposta a uma verificação de integridade do EC2 que indicou que ela tinha sido terminada ou interrompida

Problema: instâncias do Auto Scaling que foram interrompidas, reinicializadas ou terminadas são substituídas.

Causa 1: um usuário interrompeu, reinicializou ou terminou manualmente a instância.

Solução 1:: Se precisar interromper ou reinicializar as instâncias no seu grupo do Auto Scaling, recomendamos colocar as instâncias em espera primeiro. Para obter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.

Causa 2: o Amazon EC2 Auto Scaling tenta substituir instâncias spot depois que o serviço spot do Amazon EC2 interrompe as instâncias, porque o preço spot aumenta além do seu preço máximo ou a capacidade não está mais disponível.

Solução 2: não há garantia de que exista uma instância spot para atender à solicitação em qualquer momento específico. No entanto, você pode tentar o seguinte:

- Use um preço máximo spot mais alto (possivelmente, o preço sob demanda). Ao definir seu preço máximo mais alto, a chance do serviço spot do Amazon EC2 iniciar e manter a quantidade necessária de capacidade é maior.
- Aumente o número de grupos de capacidade diferentes dos quais você pode iniciar instâncias executando vários tipos de instâncias em várias zonas de disponibilidade. Para obter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.
- Se você usar vários tipos de instâncias, considere ativar o recurso de rebalanceamento de capacidade. Ele será útil se você quiser que o serviço spot do Amazon EC2 tente iniciar uma nova instância spot antes que uma instância em execução seja encerrada. Para obter mais informações, consulte Rebalanceamento de capacidade no Auto Scaling para substituir instâncias spot em risco.

Causa 3: O Amazon EC2 encerrará todas as instâncias que ainda estiverem em execução 30 minutos antes do horário final do bloco de capacidade. Essa terminação abrupta faz com que o seu grupo do Auto Scaling tente iniciar novas instâncias para manter a capacidade desejada, mesmo que o bloco de capacidade esteja terminando.

Solução 3: Para resolver esse problema, tente as seguintes soluções:

- Diminua a capacidade desejada do grupo do Auto Scaling para evitar que ele tente iniciar novas instâncias. Para obter mais informações, consulte Escalabilidade manual para o Amazon EC2 Auto Scaling.
- Certifique-se de reduzir a escala horizontalmente em seu grupo do Auto Scaling 30 minutos antes do horário de término do bloco de capacidade para que você não encontre esse erro com frequência. Certifique-se de que os ganchos do ciclo de vida tenham sido concluídos 30 minutos antes do horário de término do bloco de capacidade. Para obter mais informações, consulte Use Capacity Blocks para workloads de machine learning..

Uma instância foi retirada de serviço em resposta a uma falha na verificação de integridade do sistema ELB

Problema: instâncias do Auto Scaling poderiam ser aprovadas nas verificações de status do EC2. Mas elas poderiam falhar nas verificações de saúde do Elastic Load Balancing para os grupos de destino ou Classic Load Balancers com os quais o grupo do Auto Scaling está registrado.

Causa 1: se o seu grupo do Auto Scaling depender de verificações de integridade fornecidas pelo Elastic Load Balancing, o Amazon EC2 Auto Scaling determinará o status da integridade

de suas instâncias verificando os resultados tanto das verificações de status do EC2 quanto das verificações de integridade do Elastic Load Balancing. O balanceador de carga executa verificações de integridade enviando uma solicitação para cada instância e aguardando a resposta correta ou estabelecendo uma conexão com a instância. Uma instância pode falhar na verificação de integridade do Elastic Load Balancing porque uma aplicação em execução na instância tem problemas que fazem com que o balanceador de carga a considere fora de serviço.

Solução 1: para passar nas verificações de integridade do Elastic Load Balancing:

- Verifique se as configurações da verificação de integridade dos seus grupos de destino estão configuradas corretamente. Você define as configurações de verificação de integridade para seu balanceador de carga por grupo de destino. Para obter mais informações, consulte Configurar verificações de integridade para destinos.
- Anote os códigos de sucesso que o balanceador de carga está esperando e verifique se a aplicação está configurada corretamente para retornar esses códigos com sucesso.
- Verifique se os grupos de segurança do balanceador de carga e do grupo do Auto Scaling estão configurados corretamente.
- Verifique se o balanceador de carga está configurado nas mesmas zonas de disponibilidade do grupo do Auto Scaling.

Solução 2: atualizar o grupo do Auto Scaling para desativar as verificações de integridade do Elastic Load Balancing. Para obter instruções sobre como desativar essas verificações de integridade, consulte Desvinculação de um grupo de destino ou Classic Load Balancer.

Causa 2: há uma incompatibilidade entre o período de carência da verificação de integridade e o tempo de inicialização da instância.

Solução 3: Edite o período de carência da verificação de integridade para o grupo do Auto Scaling. Defina o período de carência da verificação de integridade do seu grupo do Auto Scaling como um período suficientemente longo para suportar o número de verificações de integridade consecutivas bem-sucedidas necessárias antes que o Elastic Load Balancing considere uma instância recém-iniciada como íntegra. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

Recursos adicionais do

Se você tiver um problema diferente, consulte os seguintes AWS re:Post artigos para obter ajuda adicional na solução de problemas:

- Por que o Amazon EC2 Auto Scaling falhou ao terminar uma instância?
- Por que o Amazon EC2 Auto Scaling não encerrou uma instância não íntegra?

Health Dashboard notificações para Amazon EC2 Auto Scaling

Você Health Dashboard fornece suporte para notificações provenientes do Amazon EC2 Auto Scaling. Essas notificações proporcionam conhecimento e orientação de remediação para problemas de performance de recursos ou disponibilidade que podem afetar suas aplicações. Somente eventos específicos para modelos de execução e grupos de segurança ausentes estão disponíveis no momento.

Isso Health Dashboard faz parte do AWS Health serviço. Ele não requer configuração e pode ser visualizado por qualquer usuário autenticado em sua conta. Para obter mais informações, consulte Introdução ao seu AWS Health painel.

Se você receber uma mensagem semelhante às seguintes, ela deverá ser tratada como um alarme para executar uma ação.

Exemplo: a escala do grupo do Auto Scaling não está aumentando na horizontal devido a um grupo de segurança ausente

```
 Hello,

 At 2020-01-11 04:00 UTC, we detected an issue with your Auto Scaling group [ARN] in
Conta da AWS 123456789012.

 A security group associated with this Auto Scaling group cannot be found. Each time
a
 scale out operation is performed, it will be prevented until you make a change that
 fixes the issue.

 We recommend that you review and update your Auto Scaling group configuration to
change
 the launch template or launch configuration that depends on the unavailable security
 group.

 Sincerely,
 Amazon Web Services
```

Exemplo: a escala do grupo do Auto Scaling não está aumentando na horizontal devido a um modelo de execução ausente

```
 Hello,

 At 2021-05-11 04:00 UTC, we detected an issue with your Auto Scaling group [ARN] in
Conta da AWS 123456789012.

 The launch template associated with this Auto Scaling group cannot be found. Each
time
 a scale out operation is performed, it will be prevented until you make a change
that
 fixes the issue.

 We recommend that you review and update your Auto Scaling group configuration and
 specify an existing launch template to use.

 Sincerely,
 Amazon Web Services
```

Monitore CloudWatch métricas para seus grupos e instâncias do Auto Scaling

As métricas são o conceito fundamental na Amazon CloudWatch. Uma métrica representa um conjunto ordenado por tempo de pontos de dados publicados em. CloudWatch Considere uma métrica como variável a ser monitorada, e os pontos de dados representando os valores dessa variável ao longo do tempo. Você pode usar essas métricas para verificar se o sistema está executando conforme o esperado.

As métricas do Amazon EC2 Auto Scaling que coletam informações sobre os grupos do Auto Scaling estão no namespace `AWS/AutoScaling`. As métricas de instância do Amazon EC2 que coletam dados de CPU e outros dados de uso de instâncias do Auto Scaling estão no namespace`AWS/EC2`.

O console do Amazon EC2 Auto Scaling exibe uma série de gráficos para as métricas do grupo e as métricas de instância agregadas para o grupo. Dependendo de suas necessidades, você pode preferir acessar dados para seus grupos e instâncias do Auto Scaling da Amazon CloudWatch em vez do console Amazon EC2 Auto Scaling.

Para obter mais informações, consulte o Guia CloudWatch do usuário da Amazon.

Conteúdo

- Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling
- CloudWatch Métricas da Amazon para Amazon EC2 Auto Scaling
- Configurar monitoramento para instâncias do Auto Scaling

Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling

Na seção Amazon EC2 Auto Scaling do console do Amazon EC2, você minute-by-minute pode monitorar o progresso de grupos individuais do Auto Scaling usando métricas. CloudWatch

É possível monitorar os seguintes tipos de métricas:

- Métricas do Auto Scaling – as métricas de Auto Scaling são ativadas somente quando você as habilita. Para obter mais informações, consulte Ativar métricas do grupo do Auto Scaling (console). Quando as métricas do Auto Scaling estão habilitadas, os gráficos de monitoramento mostram dados publicados em granularidade de um minuto para métricas de Auto Scaling.
- Métricas do EC2 – as métricas da instância do Amazon EC2 estão sempre habilitadas. Se o monitoramento detalhado estiver habilitado, os gráficos de monitoramento mostrarão dados publicados em granularidade de um minuto para métricas de instância. Para obter mais informações, consulte Configurar monitoramento para instâncias do Auto Scaling.

Para visualizar gráficos de monitoramento usando o console do Amazon EC2 Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo do Auto Scaling para o qual deseja visualizar métricas.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Escolha a guia Monitoring (Monitoramento).

   O Amazon EC2 Auto Scaling exibe gráficos de monitoramento para métricas do Auto Scaling.
4. Para visualizar gráficos de monitoramento das métricas agregadas de instância para o grupo, selecione EC2.

Ações de gráfico

- Passe o mouse sobre um ponto dos dados para exibir uma janela pop-up de dados para um horário específico em UTC.
- Para ampliar um gráfico, selecione Enlarge (Ampliar) na ferramenta de menu (os três pontos verticais) no canto superior direito do gráfico. Como alternativa, selecione o ícone de maximização na parte superior do gráfico.
- Ajuste o período de tempo para os dados exibidos no gráfico, selecionando um dos valores predefinidos do período de tempo. Se o gráfico estiver ampliado, você pode selecionar Custom (Personalizar) para definir seu próprio período de tempo.
- Selecione Refresh (Atualizar) na ferramenta de menu para atualizar os dados em um gráfico.
- Arraste o cursor sobre os dados do gráfico para selecionar um intervalo específico. Então, será possível selecionar Apply time range (Aplicar intervalo de tempo) na ferramenta de menu.
- Escolha Exibir registros na ferramenta de menu para visualizar os fluxos de registros associados (se houver) no CloudWatch console.
- Para visualizar um gráfico CloudWatch, escolha Exibir em métricas na ferramenta de menu. Isso leva você à CloudWatch página desse gráfico. Lá, você pode visualizar mais informações ou acessar informações históricas para entender melhor como seu grupo do Auto Scaling mudou ao longo de um período extenso.

Métricas de gráficos para seus grupos do Auto Scaling

Depois de criar um grupo do Auto Scaling, você poderá abrir o console do Amazon EC2 Auto Scaling e visualizar uma série de gráficos de monitoramento para o grupo na guia Monitoring (Monitoramento).

Na seção Auto Scaling, as métricas do gráfico incluem as seguintes métricas. Essas métricas fornecem medições que podem ser indicadores de um problema potencial, como número de instâncias de terminação ou número de instâncias pendentes. Você pode encontrar definições para essas métricas em CloudWatch Métricas da Amazon para Amazon EC2 Auto Scaling.

Nome de exibição	CloudWatch nome da métrica
Tamanho mínimo do grupo	`GroupMinSize`
Tamanho máximo do grupo	`GroupMaxSize`

Nome de exibição	CloudWatch nome da métrica
Capacidade desejada	`GroupDesiredCapacity`
Em instâncias de serviço	`GroupInServiceInstances`
Instâncias pendentes	`GroupPendingInstances`
Instâncias em espera	`GroupStandbyInstances`
Instâncias em encerramento	`GroupTerminatingInstances`
Total de instâncias	`GroupTotalInstances`

Na seção EC2, você pode encontrar as seguintes métricas de gráfico com base nas métricas de performance cruciais para suas instâncias do Amazon EC2. Essas métricas do EC2 são um agregado de métricas de todas as instâncias do grupo. Você pode encontrar definições para essas métricas em Listar as CloudWatch métricas disponíveis para suas instâncias no Guia do usuário do Amazon EC2.

Nome de exibição	CloudWatch nome da métrica
Utilização da CPU	`CPUUtilization`
Leituras de disco	`DiskReadBytes`
Operações de leitura de disco	`DiskReadOps`
Gravações de disco	`DiskWriteBytes`
Operações de gravação de disco	`DiskWriteOps`
Entrada de rede	`NetworkIn`
Saída de rede	`NetworkOut`
Falha na verificação de status (qualquer)	`StatusCheckFailed`

Nome de exibição	CloudWatch nome da métrica
Falha na verificação de status (instância)	`StatusCheckFailed_Instance`
Falha na verificação de status (sistema)	`StatusCheckFailed_System`

Além disso, algumas métricas estão disponíveis para casos de uso específicos nas métricas do gráfico do Auto Scaling.

As especificações dos gráficos a seguir são úteis se o seu grupo usa pesos que definem quantas unidades de instância estão disponíveis para a capacidade desejada do grupo. Você pode encontrar definições para essas métricas em CloudWatch Métricas da Amazon para Amazon EC2 Auto Scaling.

Nome de exibição	CloudWatch nome da métrica
Unidades de capacidade em serviço	`GroupInServiceCapacity`
Unidades de capacidade pendentes	`GroupPendingCapacity`
Unidades de capacidade em espera	`GroupStandbyCapacity`
Unidades de capacidade em encerramento	`GroupTerminatingCapacity`
Total de unidades de capacidade	`GroupTotalCapacity`

As métricas a seguir são úteis se o seu grupo usa o recurso de pool ativo. Você pode encontrar definições para essas métricas em CloudWatch Métricas da Amazon para Amazon EC2 Auto Scaling.

Nome de exibição	CloudWatch nome da métrica
Tamanho mínimo do pool ativoWarm Pool Minimum Size	`WarmPoolMinSize`
Capacidade desejada do pool ativo	`WarmPoolDesiredCapacity`
Unidades de capacidade pendentes no pool ativo	`WarmPoolPendingCapacity`
Unidades de capacidade em encerramento no pool ativo	`WarmPoolTerminatingCapacity`
Unidades de capacidade aquecidas no pool ativo	`WarmPoolWarmedCapacity`
Total de unidades de capacidade executadas no pool ativo	`WarmPoolTotalCapacity`
Capacidade desejada do grupo e do pool ativo	`GroupAndWarmPoolDesiredCapacity`
Total de unidades de capacidade executadas no grupo e no pool ativo	`GroupAndWarmPoolTotalCapacity`

Recursos relacionados

- Para monitorar métricas por instância, consulte Graph metrics for your instances (Métricas de gráfico para suas instâncias) no Guia do usuário do Amazon EC2 para instâncias do Linux.
- CloudWatch os painéis são páginas iniciais personalizáveis no CloudWatch console. Você pode usar essas páginas para monitorar seus recursos em uma única visualização, incluindo recursos distribuídos por regiões diferentes. Você pode usar CloudWatch painéis para criar visualizações personalizadas das métricas e alarmes de seus AWS recursos. Para obter mais informações, consulte o Guia CloudWatch do usuário da Amazon.

CloudWatch Métricas da Amazon para Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling publica as seguintes métricas no namespace`AWS/AutoScaling`. As métricas reais do grupo do Auto Scaling disponibilizadas dependem das métricas de grupo habilitadas e de quais métricas de grupo você habilitou. As métricas do grupo estão disponíveis a uma granularidade de um minuto sem custo adicional, mas é necessário habilitá-las.

Quando você ativa as métricas de grupo do Auto Scaling, o Amazon EC2 Auto Scaling envia dados CloudWatch amostrados a cada minuto, com base no melhor esforço possível. Em casos raros, quando ocorre CloudWatch uma interrupção no serviço, os dados não são preenchidos para preencher lacunas no histórico de métricas do grupo.

Conteúdo

- Métricas do grupo do Auto Scaling
- Dimensões para métricas do grupo do Auto Scaling
- Métricas e dimensões de escalabilidade preditiva
- Ativar métricas do grupo do Auto Scaling (console)
- Habilitar métricas do grupo do Auto Scaling (AWS CLI)

Métricas do grupo do Auto Scaling

Com essas métricas, você obtém visibilidade quase contínua sobre o histórico de seu grupo do Auto Scaling, como alterações no tamanho do grupo ao longo do tempo.

Métrica	Description
`GroupMinSize`	O tamanho mínimo do grupo do Auto Scaling. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupMaxSize`	O tamanho máximo do grupo do Auto Scaling. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupDesiredCapacity`	O número de instâncias que o grupo do Auto Scaling tenta manter.

Métrica	Description
	Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupInServiceInstances`	O número de instâncias que estão sendo executadas como parte do grupo do Auto Scaling. Essa métrica não inclui instâncias pendentes ou sendo encerradas. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupPendingInstances`	O número de instâncias pendentes. Uma instância pendente ainda não está em serviço. Essa métrica não inclui instâncias em serviço ou sendo encerradas. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupStandbyInstances`	O número de instâncias que estão em um estado `Standby`. As instâncias nesse estado ainda estão em execução, mas não estão ativamente em serviço. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupTerminatingInstances`	O número de instâncias que estão em processo de encerramento. Essa métrica não inclui instâncias que estejam em serviço, pendentes ou retornando a um grupo de alta atividade após a redução da escala na horizontal do grupo do Auto Scaling. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupTerminatingRetainedInstances`	O número de instâncias no `Terminating:Retained` estado. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.

Métrica	Description
`GroupTotalInstances`	O número total de instâncias no grupo do Auto Scaling. Essa métrica identifica o número de instâncias que estão em serviço, pendentes e sendo encerradas. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.

Quando você configura um grupo misto de instâncias para medir a capacidade desejada em unidades diferentes, por exemplo, atribuindo pesos com base na contagem de vCPU de cada tipo de instância, as métricas contam o número de unidades usadas pelo seu grupo do Auto Scaling. Se você não configurou um grupo de instâncias mistas para medir a capacidade desejada em unidades diferentes, as métricas seguintes são preenchidas, mas são iguais às métricas definidas na tabela anterior. Para obter mais informações, consulte Visão geral da configuração para criar um grupo de instâncias mistas.

Métrica	Description
`GroupInServiceCapacity`	O número de unidades de capacidade em execução como parte do grupo do Auto Scaling. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupPendingCapacity`	O número de unidades de capacidade pendentes. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupStandbyCapacity`	O número de unidades de capacidade que estão em um estado `Standby`. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupTerminatingCapacity`	O número de unidades de capacidade que estão em processo de encerramento.

Métrica	Description
	Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupTerminatingRetainedCapacity`	As unidades de capacidade representadas por instâncias no `Terminating:Retained` estado. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupTotalCapacity`	O número total de unidades de capacidade no grupo do Auto Scaling. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.

O Amazon EC2 Auto Scaling também relata as seguintes métricas para os grupos do Auto Scaling que têm um pool ativo. Para obter mais informações, consulte Diminua a latência para aplicações com tempos de inicialização longos usando pools ativos.

Métrica	Description
`WarmPoolMinSize`	O tamanho mínimo do pool ativo. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`WarmPoolDesiredCapacity`	A quantidade de capacidade que o Amazon EC2 Auto Scaling tenta manter no pool ativo. Isso equivale ao tamanho máximo do grupo do Auto Scaling menos a sua capacidade desejada ou, se definido, como a capacidade máxima preparada do grupo do Auto Scaling menos a sua capacidade desejada. No entanto, quando o tamanho mínimo do pool ativo for igual ou maior que a diferença entre o tamanho máximo (ou, se definido, a capacidade máxima preparada) e a capacidade desejada do

Métrica	Description
	grupo do Auto Scaling, a capacidade desejada do pool ativo será equivalente a `WarmPoolMinSize` . Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`WarmPoolPendingCapacity`	A quantidade de capacidade no pool ativo que está pendente. Isso inclui instâncias que retornam a um grupo de alta atividade após a redução da escala na horizontal do grupo do Auto Scaling. Essa métrica não inclui instâncias em execução, interrompidas ou sendo terminadas. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`WarmPoolPendingRetainedCapacity`	Rastreia instâncias retidas retornando à piscina aquecida. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`WarmPoolTerminatingCapacity`	A quantidade de capacidade no pool ativo que está em processo de encerramento. Essa métrica não inclui instâncias em execução, interrompidas ou pendentes. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`WarmPoolTerminatingRetainedCapacity`	Rastreia instâncias retidas que terminam na piscina aquecida. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`WarmPoolWarmedCapacity`	A quantidade de capacidade disponível para se inserir no grupo do Auto Scaling durante a redução da escala. Essa métrica não inclui instâncias pendentes ou sendo encerradas. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.

Métrica	Description
`WarmPoolTotalCapacity`	A capacidade total do pool ativo, incluindo instâncias que estão em execução, interrompidas, pendentes ou sendo terminadas. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupAndWarmPoolDesiredCapacity`	A capacidade desejada do grupo do Auto Scaling e o pool ativo combinados. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`GroupAndWarmPoolTotalCapacity`	A capacidade total do grupo do Auto Scaling e o pool ativo combinados. Isso inclui instâncias que estão sendo em execução, interrompidas, pendentes, sendo terminadas ou em serviço. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.

Dimensões para métricas do grupo do Auto Scaling

É possível usar as seguintes dimensões para refinar as métricas listadas nas tabelas anteriores.

Dimensão	Description
`AutoScalingGroupName`	Filtros no nome de um grupo do Auto Scaling.

Métricas e dimensões de escalabilidade preditiva

O namespace `AWS/AutoScaling` inclui as métricas a seguir para escalabilidade preditiva.

As métricas estão disponíveis com uma resolução de uma hora.

Você pode avaliar a precisão da previsão comparando os valores previstos com os valores efetivos. Para obter mais informações sobre como avaliar a precisão da previsão usando essas métricas, consulte Monitore métricas de escalabilidade preditiva com CloudWatch.

Métrica	Description	Dimensões
`PredictiveScalingLoadForecast`	A previsão da quantidade de carga que será gerada por sua aplicação. As estatísticas `Average`, `Minimum` e `Maximum` são úteis, mas a estatística `Sum` não. Critérios de relatório: reportado após a criação da previsão inicial.	`AutoScalingGroupName`, `PolicyName`, `PairIndex`
`PredictiveScalingCapacityForecast`	A quantidade prevista de capacidade necessária para atender à demanda de aplicações. Isso se baseia na previsão de carga e no nível de utilização pretendido no qual você deseja manter suas instâncias do Auto Scaling. As estatísticas`Average`, `Minimum` e `Maximum` são úteis, mas a estatística `Sum` não. Critérios de relatório: reportado após a criação da previsão inicial.	`AutoScalingGroupName`, `PolicyName`
`PredictiveScalingMetricPairCorrelation`	A correlação entre a métrica de escalabilidade e a média por instância da métrica de carga. A escalabilidade preditiva pressupõe alta correlação. Então, se você observar um valor baixo para essa métrica, é melhor não usar um par de métricas. As estatísticas`Average`, `Minimum` e `Maximum` são úteis, mas a estatística `Sum` não. Critérios de relatório: reportado após a criação da previsão inicial.	`AutoScalingGroupName`, `PolicyName`, `PairIndex`

> **Note**
>
> A dimensão `PairIndex` retorna informações associadas ao índice do par de métricas de escalabilidade de carga, conforme atribuído pelo Amazon EC2 Auto Scaling. Atualmente, o único valor válido é `0`.

Ativar métricas do grupo do Auto Scaling (console)

Para habilitar as métricas do grupo

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Monitoring (Monitoramento), marque a caixa de seleção Enable (Habilitar) em Auto Scaling group metrics collection (Coleta de métricas do grupo do Auto Scaling) localizada na parte superior da página em Auto Scaling.

Para desabilitar as métricas do grupo

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Selecione seu grupo do Auto Scaling.
3. Na guia Monitoring (Monitoramento), em Auto Scaling group metrics collection (Coleta de métricas do grupo do Auto Scaling), desmarque a caixa de seleção Enable (Habilitar).

Habilitar métricas do grupo do Auto Scaling (AWS CLI)

Para habilitar métricas do grupo do Auto Scaling

Ative uma ou mais métricas de grupo usando o enable-metrics-collectioncomando. Por exemplo, o comando a seguir habilita uma única métrica para o grupo do Auto Scaling especificado.

```
aws autoscaling enable-metrics-collection --auto-scaling-group-name my-asg \
  --metrics GroupDesiredCapacity --granularity "1Minute"
```

Se você omitir a opção`--metrics`, todas as métricas serão habilitadas.

```
aws autoscaling enable-metrics-collection --auto-scaling-group-name my-asg \
  --granularity "1Minute"
```

Para desabilitar métricas do grupo do Auto Scaling

Use o disable-metrics-collectioncomando para desativar todas as métricas do grupo.

```
aws autoscaling disable-metrics-collection --auto-scaling-group-name my-asg
```

Configurar monitoramento para instâncias do Auto Scaling

O Amazon EC2 coleta e processa os dados brutos das instâncias, e os transforma em métricas legíveis e praticamente em tempo real que descrevem o uso de CPU e outros dados de uso do grupo do Auto Scaling. Você pode configurar o intervalo para monitorar essas métricas escolhendo a granularidade de um ou cinco minutos.

Sempre que uma instância for executada, o monitoramento será habilitado usando monitoramento básico (granularidade de cinco minutos) ou monitoramento detalhado (granularidade de um minuto). Para o monitoramento detalhado, aplicam-se custos adicionais. Para obter mais informações, consulte os CloudWatch preços da Amazon e o monitoramento de suas instâncias usando CloudWatch no Guia do usuário do Amazon EC2.

Para criar um grupo do Auto Scaling, é necessário criar uma configuração de execução ou um modelo de execução que permita o tipo de monitoramento adequado à sua aplicação. Se você adicionar uma política de escalabilidade ao seu grupo, é altamente recomendável usar o monitoramento detalhado para obter dados de métricas para instâncias do EC2 com uma granularidade de um minuto, pois isso atingirá uma resposta mais rápida a alterações na carga.

Conteúdo

- Habilitar o monitoramento detalhado (console)
- Habilitar o monitoramento detalhado (AWS CLI)
- Alternar entre monitoramento básico e detalhado
- Colete métricas adicionais usando o CloudWatch agente

Habilitar o monitoramento detalhado (console)

Por padrão, o monitoramento básico é ativado quando você usa o Console de gerenciamento da AWS para criar um modelo de execução ou configuração de execução.

Para habilitar o monitoramento detalhado em um modelo de execução

Ao criar o modelo de lançamento usando o Console de gerenciamento da AWS, na seção Detalhes avançados, para CloudWatchMonitoramento detalhado, escolha Habilitar. Caso contrário, o monitoramento básico será habilitado. Para obter mais informações, consulte Criar um modelo de execução usando configurações avançadas.

Para habilitar o monitoramento detalhado em uma configuração de execução

Ao criar a configuração de execução usando o Console de gerenciamento da AWS, na seção Configuração adicional, selecione Habilitar monitoramento detalhado da instância EC2 em CloudWatch. Caso contrário, o monitoramento básico será habilitado. Para obter mais informações, consulte Criar uma configuração de execução.

Habilitar o monitoramento detalhado (AWS CLI)

Por padrão, o monitoramento básico é habilitado quando você cria um modelo de execução usando a AWS CLI. O monitoramento detalhado é habilitado por padrão quando você cria uma configuração de execução usando a AWS CLI ou um SDK.

Para habilitar o monitoramento detalhado em um modelo de execução

Para modelos de execução, use o comando create-launch-template e envie um arquivo JSON que contenha as informações para criar o modelo de execução. Defina o parâmetro de monitoramento como `"Monitoring":{"Enabled":true}` para habilitar o monitoramento detalhado ou `"Monitoring":{"Enabled":false}` para habilitar o monitoramento básico.

Para habilitar o monitoramento detalhado em uma configuração de execução

Para configurações de execução, use o comando create-launch-configuration com a opção `--instance-monitoring`. Defina essa opção como `true` para habilitar o monitoramento detalhado ou `false` para habilitar o monitoramento básico.

```
--instance-monitoring Enabled=true
```

Alternar entre monitoramento básico e detalhado

Para alterar o tipo de monitoramento habilitado em novas instâncias do EC2, atualize o modelo de execução ou o grupo do Auto Scaling para usar um novo modelo ou uma nova configuração de execução. As instâncias existentes continuam a usar o tipo de monitoramento habilitado anteriormente. Para atualizar todas as instâncias, termine-as para que elas sejam substituídas por seu grupo do Auto Scaling ou atualize as instâncias individualmente usando monitor-instances e unmonitor-instances.

> **Note**
>
> Com os recursos de tempo de vida máximo e atualização de instância e de atualização da instância, também é possível substituir todas as instâncias no grupo do Auto Scaling para iniciar novas instâncias que usem as novas configurações. Para obter mais informações, consulte Recicle as instâncias em seu grupo do Auto Scaling.

Ao alternar entre monitoramento básico e detalhado:

Se você tiver CloudWatch alarmes associados às políticas de escalonamento de etapas ou políticas de escalabilidade simples para seu grupo de Auto Scaling, use o put-metric-alarmcomando para atualizar cada alarme. Faça com que cada período corresponda ao tipo de monitoramento (300 segundos para o monitoramento básico e 60 segundos para o monitoramento detalhado). Se você passar do monitoramento detalhado para o monitoramento básico, mas não atualizar seus alarmes para corresponderem ao período de cinco minutos, eles continuarão a verificar as estatísticas a cada minuto. Eles poderão não encontrar nenhum dado disponível para quatro de cada cinco períodos.

Colete métricas adicionais usando o CloudWatch agente

Para coletar métricas no nível do sistema operacional, como memória disponível e usada, você deve instalar o CloudWatch agente. Podem ser cobrados taxas adicionais. Você pode usar o CloudWatch agente para coletar métricas do sistema e arquivos de log das instâncias do Amazon EC2. Para obter mais informações, consulte Métricas coletadas pelo CloudWatch agente no Guia CloudWatch do usuário da Amazon.

Registre chamadas da API Amazon EC2 Auto Scaling com AWS CloudTrail

O Amazon EC2 Auto Scaling é AWS CloudTrailintegrado a um serviço que fornece um registro das ações realizadas por um usuário, função ou um. AWS service (Serviço da AWS) CloudTrail captura chamadas de API para Auto Scaling como eventos. As chamadas capturadas incluem as chamadas do Console de gerenciamento da AWS e as chamadas de código à API de operações do Application Auto Scaling. Usando as informações coletadas por CloudTrail, você pode determinar a solicitação que foi feita ao Auto Scaling, o endereço IP a partir do qual a solicitação foi feita, quando foi feita e detalhes adicionais.

Cada entrada de log ou evento contém informações sobre quem gerou a solicitação. As informações de identidade ajudam a determinar o seguinte:

- Se a solicitação foi feita com credenciais de usuário raiz ou credenciais de usuário.
- Se a solicitação foi feita em nome de um usuário do Centro de Identidade do IAM.
- Se a solicitação foi feita com credenciais de segurança temporárias de um perfil ou de um usuário federado.
- Se a solicitação foi feita por outro AWS service (Serviço da AWS).

CloudTrail está ativo Conta da AWS quando você cria a conta e você tem acesso automático ao histórico de CloudTrail eventos. O histórico de CloudTrail eventos fornece um registro visível, pesquisável, baixável e imutável dos últimos 90 dias de eventos de gerenciamento registrados em um. Região da AWS Para obter mais informações, consulte Trabalhando com o histórico de CloudTrail eventos no Guia AWS CloudTrail do usuário. Não há CloudTrail cobrança pela visualização do histórico de eventos.

Para um registro contínuo dos eventos dos Conta da AWS últimos 90 dias, crie uma trilha.

CloudTrail trilhas

Uma trilha permite CloudTrail entregar arquivos de log para um bucket do Amazon S3. Todas as trilhas criadas usando o Console de gerenciamento da AWS são multirregionais. Só é possível criar uma trilha de região única ou de várias regiões usando a AWS CLI. É recomendável criar uma trilha multirregional porque você captura todas as atividades Regiões da AWS em sua conta. Ao criar uma trilha de região única, é possível visualizar somente os eventos registrados na Região da AWS da trilha. Para obter mais informações sobre trilhas, consulte Criar uma trilha

para a Conta da AWS e Criar uma trilha para uma organização no Guia do usuário do AWS CloudTrail .

Você pode entregar uma cópia dos seus eventos de gerenciamento contínuos para o bucket do Amazon S3 sem nenhum custo CloudTrail criando uma trilha. No entanto, há cobranças de armazenamento do Amazon S3. Para obter mais informações sobre CloudTrail preços, consulte AWS CloudTrail Preços. Para receber informações sobre a definição de preços do Amazon S3, consulte Definição de preços do Amazon S3.

Eventos de gerenciamento do Auto Scaling em CloudTrail

Os eventos de gerenciamento fornecem informações sobre as operações de gerenciamento que são realizadas nos recursos do seu Conta da AWS. Também são conhecidas como operações de ambiente de gerenciamento. Por padrão, CloudTrail registra eventos de gerenciamento.

O Amazon EC2 Auto Scaling registra em log todas as operações do ambiente de gerenciamento do Auto Scaling como eventos de gerenciamento. Para obter uma lista das operações do plano de controle do Amazon EC2 Auto Scaling nas quais o Auto Scaling se CloudTrail registra, consulte a Referência da API do Amazon EC2 Auto Scaling.

Exemplos do Auto Scaling usando o

Um evento representa uma única solicitação de qualquer fonte e inclui informações sobre a operação de API solicitada, a data e a hora da operação, os parâmetros da solicitação e assim por diante. CloudTrail os arquivos de log não são um rastreamento de pilha ordenado das chamadas públicas de API, portanto, os eventos não aparecem em nenhuma ordem específica.

O exemplo a seguir mostra um CloudTrail evento que demonstra a `CreateLaunchConfiguration` operação.


```
{
    "eventVersion": "1.05",
    "userIdentity": {
        "type": "Root",
        "principalId": "123456789012",
        "arn": "arn:aws:iam::123456789012:root",
        "accountId": "123456789012",
        "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
        "sessionContext": {
            "attributes": {
```

```
                "mfaAuthenticated": "false",
                "creationDate": "2018-08-21T17:05:42Z"
            }
        }
    },
    "eventTime": "2018-08-21T17:07:49Z",
    "eventSource": "autoscaling.amazonaws.com",
    "eventName": "CreateLaunchConfiguration",
    "awsRegion": "us-west-2",
    "sourceIPAddress": "192.0.2.0",
    "userAgent": "Coral/Jakarta",
    "requestParameters": {
        "ebsOptimized": false,
        "instanceMonitoring": {
            "enabled": false
        },
        "instanceType": "t2.micro",
        "keyName": "EC2-key-pair-oregon",
        "blockDeviceMappings": [
            {
                "deviceName": "/dev/xvda",
                "ebs": {
                    "deleteOnTermination": true,
                    "volumeSize": 8,
                    "snapshotId": "snap-01676e0a2c3c7de9e",
                    "volumeType": "gp2"
                }
            }
        ],
        "launchConfigurationName": "launch_configuration_1",
        "imageId": "ami-6cd6f714d79675a5",
        "securityGroups": [
            "sg-00c429965fd921483"
        ]
    },
    "responseElements": null,
    "requestID": "0737e2ea-fb2d-11e3-bfd8-99133058e7bb",
    "eventID": "3fcfb182-98f8-4744-bd45-b38835ab61cb",
    "eventType": "AwsApiCall",
    "recipientAccountId": "123456789012"
}
```

Para obter informações sobre o conteúdo do CloudTrail registro, consulte o conteúdo do CloudTrail registro no Guia AWS CloudTrail do usuário.

Chamadas de Auto Scaling ativadas RemoveAction CloudWatch

Seu AWS CloudTrail registro pode mostrar que o Auto Scaling chama a CloudWatch `RemoveAction` API quando o Auto Scaling CloudWatch instrui a remover a ação de escalonamento automático de um alarme. Isso pode acontecer se você cancelar o registro de um destino escalável, excluir uma política de escalabilidade ou se um alarme invocar uma política de escalabilidade inexistente.

Configurar notificações do Amazon SNS para o Amazon EC2 Auto Scaling

Você pode configurar seu grupo do Auto Scaling para que ele informe sobre eventos importantes que afetam sua aplicação. Com as notificações, você também pode eliminar a sondagem e não encontrará o erro `RequestLimitExceeded` que às vezes resulta da sondagem.

Há duas maneiras de receber notificações sobre o Amazon EC2 Auto Scaling:

- Amazon Simple Notification Service: Amazon SNS pode notificar você quando seu grupo do Auto Scaling iniciar ou terminar instâncias. Só é possível ativar e desativar notificações do Amazon SNS. Para obter mais informações, consulte Amazon SNS e Amazon EC2 Auto Scaling.
- Amazon EventBridge — EventBridge fornece notificações mais avançadas, orientadas por eventos, de acordo com critérios específicos e enviadas para uma variedade de destinos, incluindo o Amazon SNS. EventBridge também pode monitorar uma variedade maior de eventos do Auto Scaling para um monitoramento mais preciso. Para obter mais informações, consulte Use EventBridge para lidar com eventos do Auto Scaling.

Opcionalmente, você pode usar notificações com ganchos do ciclo de vida para executar ações personalizadas em instâncias durante a execução ou terminação. Para obter mais informações sobre como configurar as notificações a serem usadas com ganchos do ciclo de vida, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

Amazon SNS e Amazon EC2 Auto Scaling

Esta seção mostra como usar o Amazon SNS para monitorar quando o seu grupo do Auto Scaling executa ou termina instâncias.

Por exemplo, se você configurar o grupo do Auto Scaling para usar o tipo de notificação `autoscaling: EC2_INSTANCE_TERMINATE` e seu grupo do Auto Scaling terminar uma instância, ele enviará uma notificação por e-mail. Esse e-mail contém os detalhes da instância encerrada, como o ID da instância e o motivo pelo qual a instância foi encerrada.

Observe que, à medida que o Amazon EC2 Auto Scaling adiciona ou remove instâncias do grupo, as notificações sobre essas alterações são enviadas a você, com uma notificação enviada por instância. No entanto, a entrega dessas notificações é feita com base no melhor esforço, e suas instâncias ainda podem falhar após a notificação inicial, por exemplo, se uma verificação de integridade posterior falhar. Para obter mais informações sobre essas verificações de integridade, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Para obter mais informações sobre tópicos do Amazon SNS, consulte o Guia do desenvolvedor do Amazon Simple Notification Service.

Conteúdo

- Notificações do SNS
- Configurar notificações do Amazon SNS para o Amazon EC2 Auto Scaling
  - Crie um tópico do Amazon SNS
  - Assinar o tópico do Amazon SNS
  - Confirmar sua assinatura do Amazon SNS
  - Configurar o grupo do Auto Scaling para enviar notificações
  - Testar a notificação
  - Excluir a configuração da notificação
- Política de chaves para um tópico do Amazon SNS criptografado

Notificações do SNS

O Amazon EC2 Auto Scaling oferece suporte ao envio de notificações do Amazon SNS quando os seguintes eventos ocorrem.

Event	Description
`autoscaling:EC2_INSTANCE_LAUNCH`	Ativação de instância bem-sucedida

Event	Description
`autoscaling:EC2_INSTANCE_LAUNCH_ERROR`	Falha na ativação da instância
`autoscaling:EC2_INSTANCE_TERMINATE`	Encerramento da instância bem-sucedido
`autoscaling:EC2_INSTANCE_TERMINATE_ERROR`	Falha no encerramento da instância

A mensagem inclui as seguintes informações:

- `Event` — O evento.
- `AccountId`: o ID da conta do Amazon Web Services.
- `AutoScalingGroupName`: o nome do grupo do Auto Scaling.
- `AutoScalingGroupARN`: o ARN do grupo do Auto Scaling.
- `EC2InstanceId` — A ID da instância do EC2.

Por exemplo:

```
Service: AWS Auto Scaling
Time: 2016-09-30T19:00:36.414Z
RequestId: 4e6156f4-a9e2-4bda-a7fd-33f2ae528958
Event: autoscaling:EC2_INSTANCE_LAUNCH
AccountId: 123456789012
AutoScalingGroupName: my-asg
AutoScalingGroupARN: arn:aws:autoscaling:region:123456789012:autoScalingGroup...
ActivityId: 4e6156f4-a9e2-4bda-a7fd-33f2ae528958
Description: Launching a new EC2 instance: i-0598c7d356eba48d7
Cause: At 2016-09-30T18:59:38Z a user request update of AutoScalingGroup constraints
 to ...
StartTime: 2016-09-30T19:00:04.445Z
EndTime: 2016-09-30T19:00:36.414Z
StatusCode: InProgress
StatusMessage:
Progress: 50
EC2InstanceId: i-0598c7d356eba48d7
Details: {"Subnet ID":"subnet-id","Availability Zone":"zone"}
```

```
Origin: AutoScalingGroup
Destination: EC2
```

Configurar notificações do Amazon SNS para o Amazon EC2 Auto Scaling

Para usar o Amazon SNS para enviar notificações por e-mail, você deve primeiro criar um tópico e, em seguida, assinar seus endereços de e-mail para o tópico.

Crie um tópico do Amazon SNS

Um tópico do SNS é um ponto de acesso lógico, um canal de comunicação que seu grupo do Auto Scaling usa para enviar notificações. Você cria um tópico especificando um nome para o tópico.

Quando você cria o nome de um tópico, ele deve atender aos seguintes requisitos:

- Ter entre 1 e 256 caracteres
- Conter letras maiúsculas e minúsculas ASCIIs, números, sublinhados ou hífens

Para obter mais informações, consulte Criação de um tópico do Amazon SNS no Guia do desenvolvedor do Amazon Simple Notification Service.

Assinar o tópico do Amazon SNS

Para receber as notificações que seu grupo do Auto Scaling envia ao tópico, você deve assinar um endpoint para o tópico. Neste procedimento, em Endpoint, especifique o endereço de e-mail no qual você deseja receber as notificações do Amazon EC2 Auto Scaling.

Para obter instruções, consulte Assinatura de um tópico do Amazon SNS no Guia do desenvolvedor do Amazon Simple Notification Service.

Confirmar sua assinatura do Amazon SNS

O Amazon SNS envia um e-mail de confirmação para o endereço de e-mail especificado na etapa anterior.

Certifique-se de abrir o e-mail em AWS Notifications (Notificações) e escolher o link para confirmar a assinatura antes de prosseguir para a próxima etapa.

Você receberá uma mensagem de confirmação de. AWS O Amazon SNS agora está configurado para receber notificações e enviar a notificação como um e-mail para o endereço de e-mail que você especificou.

Configurar o grupo do Auto Scaling para enviar notificações

Você pode configurar seu grupo do Auto Scaling para enviar notificações para o Amazon SNS quando um evento de escalabilidade ocorre, como lançamento ou término de instâncias. O Amazon SNS envia uma notificação com informações sobre as instâncias para o endereço de e-mail que você especificou.

Para configurar notificações do Amazon SNS para o seu grupo do Auto Scaling (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página, mostrando informações sobre o grupo selecionado.
3. Na guia Activity (Atividade), escolha Activity notifications (Notificações de atividades), Create notification (Criar notificação).
4. No painel Criar notificações, faça o seguinte:

   a. Em SNS Topic (Tópico do SNS), selecione o tópico do SNS.

   b. Em Event types (Tipos de eventos), selecione os eventos sobre os quais deseja enviar notificações.

   c. Escolha Criar.

Para configurar notificações do Amazon SNS para o seu grupo do Auto Scaling (AWS CLI)

Use o seguinte comando put-notification-configuration:

```
aws autoscaling put-notification-configuration --auto-scaling-group-name my-
asg --topic-arn arn --notification-types "autoscaling:EC2_INSTANCE_LAUNCH"
 "autoscaling:EC2_INSTANCE_TERMINATE"
```

Testar a notificação

Para gerar uma notificação para um evento de lançamento, atualize o grupo do Auto Scaling aumentando a capacidade desejada do grupo do Auto Scaling em 1. Você recebe uma notificação dentro de alguns minutos após a execução da instância.

Para alterar a capacidade desejada (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página Grupos do Auto Scaling mostrando informações sobre o grupo selecionado.
3. Na guia Detalhes, escolha Detalhes do grupo, Editar.
4. Em Desired capacity (Capacidade desejada), aumente o valor atual em 1. Se esse valor exceder Maximum capacity (Capacidade máxima), também será necessário aumentar o valor Maximum capacity (Capacidade máxima) em 1.
5. Selecione Atualizar.
6. Depois de alguns minutos, você receberá uma notificação para o evento. Se não for necessário ter a instância adicional executada para este teste, será possível reduzir Desired capacity (Capacidade desejada) em 1. Depois de alguns minutos, você receberá uma notificação para o evento.

Excluir a configuração da notificação

Você poderá excluir sua configuração de notificação do Amazon EC2 Auto Scaling se ela não estiver mais sendo usada.

Para excluir a configuração de notificação do Amazon EC2 Auto Scaling (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Selecione seu grupo do Auto Scaling.
3. Na guia Activity (Atividade), marque a caixa de seleção ao lado da notificação que deseja excluir e escolha Actions (Ações), Delete (Excluir).

Para excluir a configuração de notificação do Amazon EC2 Auto Scaling (AWS CLI)

Use o seguinte comando delete-notification-configuration:

```
aws autoscaling delete-notification-configuration --auto-scaling-group-name my-asg --
topic-arn arn
```

Para obter informações sobre como excluir o tópico do Amazon SNS e todas as assinaturas associadas ao seu grupo do Auto Scaling, consulte Exclusão de assinaturas e tópicos do Amazon SNS no Guia do desenvolvedor do Amazon Simple Notification Service.

Política de chaves para um tópico do Amazon SNS criptografado

O tópico do Amazon SNS que você especificar pode ser criptografado com uma chave gerenciada pelo cliente criada com o AWS Key Management Service. Para dar permissão ao Amazon EC2 Auto Scaling para publicar em tópicos criptografados, você deve primeiro criar sua chave KMS e depois adicionar a seguinte declaração à política da chave KMS. Substitua o ARN de exemplo pelo ARN da função apropriada vinculada ao serviço que tem permissão de acesso à chave. Para obter mais informações, consulte Configurar AWS KMS permissões no Guia do desenvolvedor do Amazon Simple Notification Service.

Neste exemplo, a declaração de política dá à função vinculada ao serviço denominada AWSServiceRoleForAutoScalingpermissões para usar a chave gerenciada pelo cliente. Para saber mais sobre o perfil vinculado ao serviço do Amazon EC2 Auto Scaling, consulte Perfis vinculados ao serviço do Amazon EC2 Auto Scaling.

```
{
  "Sid": "Allow service-linked role use of the customer managed key",
  "Effect": "Allow",
  "Principal": {
    "AWS": "arn:aws:iam::123456789012:role/aws-service-role/autoscaling.amazonaws.com/
AWSServiceRoleForAutoScaling"
  },
  "Action": [
    "kms:GenerateDataKey*",
    "kms:Decrypt"
  ],
  "Resource": "*"
}
```

As chaves de condição de `aws:SourceArn` e `aws:SourceAccount` não são suportadas em políticas de chaves que permitem que o Amazon EC2 Auto Scaling publique em tópicos criptografados.

AWS serviços integrados ao Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling pode ser integrado AWS a outros serviços. Veja as opções de integração a seguir para saber mais sobre como cada serviço funciona com o Amazon EC2 Auto Scaling.

Tópicos

- Rebalanceamento de capacidade no Auto Scaling para substituir instâncias spot em risco
- Reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas
- Crie grupos de Auto Scaling a partir da linha de comando usando AWS CloudShell
- Crie grupos de Auto Scaling com AWS CloudFormation
- Obtenha recomendações de tipo de instância com AWS Compute Optimizer
- Use o Elastic Load Balancing para distribuir automaticamente o tráfego de entrada de aplicações no seu grupo do Auto Scaling.
- Gerenciar o fluxo de tráfego com um grupo de destino do VPC Lattice
- Use EventBridge para lidar com eventos do Auto Scaling
- Fornecer conectividade de rede para suas instâncias do Auto Scaling usando a Amazon VPC

Rebalanceamento de capacidade no Auto Scaling para substituir instâncias spot em risco

O rebalanceamento de capacidade no Auto Scaling ajuda você a manter a disponibilidade da workload ao substituir proativamente as instâncias spot com risco de interrupção.

Quando as instâncias spot apresentam risco elevado de interrupção, o serviço do Amazon EC2 Spot envia uma recomendação de rebalanceamento do EC2 ao Amazon EC2 Auto Scaling. Se você habilitar o rebalanceamento de capacidade, o Auto Scaling tentará substituir proativamente as instâncias spot no seu grupo que tenham recebido uma recomendação de rebalanceamento de instância do EC2. Você pode decidir rebalancear sua workload em instâncias spot novas ou existentes que não tenham risco elevado de interrupção.

Quando você não usa o rebalanceamento de capacidade, o Auto Scaling não substituirá as instâncias spot até que o serviço spot do Amazon EC2 interrompa as instâncias e sua verificação

de integridade falhe. Antes de interromper uma instância, o Amazon EC2 sempre fornece uma recomendação de rebalanceamento de instância do EC2 e um aviso de interrupção de instância spot de dois minutos.

Conteúdo

- Visão geral do
- Comportamento de rebalanceamento de capacidade
- Considerações
- Habilitação do rebalanceamento de capacidade para substituir proativamente as instâncias spot em risco

Visão geral do

Para usar o rebalanceamento de capacidade com seu grupo do Auto Scaling, as etapas básicas são:

1. Configure seu grupo do Auto Scaling para usar vários tipos de instância e zonas de disponibilidade. Dessa forma, o Amazon EC2 Auto Scaling pode analisar a capacidade disponível para instâncias spot em cada zona de disponibilidade. Para obter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.
2. Adicione ganchos do ciclo de vida conforme necessário para realizar um desligamento suave da sua aplicação dentro das instâncias que recebem a notificação de rebalanceamento. Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

   A seguir estão alguns motivos pelos quais você pode usar um gancho do ciclo de vida:

   - Para o encerramento suave de operadores do Amazon SQS
   - Para concluir o cancelamento do registro do sistema de nomes de domínio (DNS)
   - Para extrair logs do sistema ou da aplicação e carregá-los no Amazon Simple Storage Service (Amazon S3)
3. Desenvolva uma ação personalizada para o gancho do ciclo de vida. Para invocar sua ação personalizada o mais rápido possível, você precisa saber quando uma instância está pronta para ser encerrada. Descubra isso detectando o estado do ciclo de vida da instância.
   - Para invocar uma ação fora da instância, escreva uma EventBridge regra e automatize a ação a ser tomada quando um padrão de evento corresponder à regra.
   - Para invocar uma ação dentro da instância, configure a instância para executar um script de desligamento e recuperar o estado do ciclo de vida por meio dos metadados da instância.

É fundamental projetar a ação personalizada para ser concluída em menos de dois minutos. Isso garante que haja tempo suficiente para concluir as tarefas antes do encerramento da instância.

Depois de concluir essas etapas, será possível começar a usar o rebalanceamento de capacidade.

Comportamento de rebalanceamento de capacidade

Com o rebalanceamento de capacidade, o Amazon EC2 Auto Scaling se comporta da seguinte maneira quando uma instância recebe uma recomendação de rebalanceamento:

- Quando a nova instância spot é iniciada, o Amazon EC2 Auto Scaling aguarda até que a nova instância passe na verificação de integridade antes de encerrar a instância anterior. Ao substituir mais de uma instância, o término de cada instância anterior é iniciado depois que a nova instância foi iniciada e aprovada na verificação de integridade.
- Como o Amazon EC2 Auto Scaling tenta iniciar novas instâncias antes de terminar as anteriores, estar na capacidade máxima especificada ou próximo a ela pode impedir ou parar completamente as atividades de rebalanceamento. Para evitar esse problema, o Amazon EC2 Auto Scaling pode exceder temporariamente o tamanho máximo do grupo em até 10% da capacidade desejada.
- Se você não adicionou um gancho do ciclo de vida ao grupo do Auto Scaling, o Amazon EC2 Auto Scaling começará a encerrar as instâncias anteriores assim que as novas instâncias passarem na verificação de integridade.
- Se você adicionou um gancho do ciclo de vida, isso aumenta o tempo necessário até começarmos a encerrar as instâncias anteriores pelo valor de tempo limite especificado para o gancho do ciclo de vida.
- Se você estiver usando políticas de escalabilidade ou escalabilidade agendada, as atividades de escalabilidade serão executadas em paralelo. Se uma ação de escalabilidade estiver em andamento e seu grupo do Auto Scaling estiver abaixo da nova capacidade desejada, o Amazon EC2 Auto Scaling aumentará a escala na horizontal antes de terminar as instâncias anteriores.

Se não houver capacidade para seus tipos de instância em uma zona de disponibilidade, o Amazon EC2 Auto Scaling continuará tentando iniciar instâncias spot em outras zonas de disponibilidade habilitadas até que ela seja bem-sucedida.

Na pior das hipóteses, se as novas instâncias falharem na inicialização ou se suas verificações de integridade falharem, o Amazon EC2 Auto Scaling continuará tentando reiniciá-las. Enquanto ele

tenta iniciar novas instâncias, as anteriores serão eventualmente interrompidas e encerradas à força com um aviso de interrupção de dois minutos.

Considerações

Considere o seguinte ao usar o reequilíbrio de capacidade:

Projete sua aplicação para ser tolerante a interrupções pontuais

Sua aplicação deve ser capaz de lidar com alterações dinâmicas no número de instâncias e com a possibilidade de uma instância spot ser interrompida antecipadamente. Por exemplo, se o seu grupo do Auto Scaling estiver atrás de um balanceador de carga do Elastic Load Balancing, o Amazon EC2 Auto Scaling aguardará o cancelamento do registro da instância no balanceador de carga antes de chamar o gancho do ciclo de vida. Se o tempo para cancelar o registro da instância e concluir a ação do ciclo de vida demorar muito, a instância poderá ser interrompida enquanto o Amazon EC2 Auto Scaling aguarda a conclusão da ação do ciclo de vida antes de encerrar a instância.

Nem sempre é possível para o Amazon EC2 enviar o sinal de recomendação de rebalanceamento antes do aviso de interrupção da instância spot de dois minutos. Às vezes, o sinal de recomendação de reequilíbrio chega ao mesmo tempo que o aviso de interrupção de dois minutos. Quando isso acontece, o Amazon EC2 Auto Scaling chama o gancho do ciclo de vida e tenta iniciar uma nova instância spot imediatamente.

Evitar um risco elevado de interrupção das instâncias spot substitutas

Suas instâncias spot de substituição poderão correr um risco elevado de interrupção se você usar a estratégia de alocação`lowest-price`. Isso ocorre porque lançamos instâncias no pool de menor preço que tem capacidade disponível naquele momento, mesmo que suas instâncias spot substitutas possam ser interrompidas logo após serem lançadas. Para evitar um alto risco de interrupção, recomendamos fortemente que não use a estratégia de alocação`lowest-price`. Em vez disso, recomendamos a estratégia de alocação`price-capacity-optimized`. Essa estratégia executa instâncias spot de substituição em grupos spot com menor probabilidade de interrupção e com o preço mais baixo possível. Portanto, é menos provável que sejam interrompidos em um futuro próximo.

O Amazon EC2 Auto Scaling só lançará uma nova instância se a disponibilidade for igual ou melhor

Um dos objetivos do rebalanceamento de capacidade é melhorar a disponibilidade de uma instância spot. Se uma instância spot existente receber uma recomendação de rebalanceamento,

o Amazon EC2 Auto Scaling só iniciará uma nova instância se a nova instância fornecer a mesma ou melhor disponibilidade do que a instância existente. Se o risco de interrupção de uma nova instância for pior do que a instância existente, o Amazon EC2 Auto Scaling não iniciará uma nova instância. No entanto, o Amazon EC2 Auto Scaling continuará avaliando os pools de capacidade Spot com base nas informações fornecidas pelo serviço Amazon EC2 Spot e iniciará uma nova instância se a disponibilidade melhorar.

Há uma chance de que sua instância existente seja interrompida sem que o Amazon EC2 Auto Scaling inicie proativamente uma nova instância. Quando isso acontece, o Amazon EC2 Auto Scaling tenta iniciar uma nova instância assim que recebe o aviso de interrupção da instância spot. Isso acontece independentemente de a nova instância ter um alto risco de interrupção.

O rebalanceamento da capacidade não aumenta a taxa de interrupção de instâncias spot

Quando o Rebalanceamento da capacidade é habilitado, ele não aumenta a Taxa de interrupção de instâncias spot (o número de instâncias spot que são recuperadas quando o Amazon EC2 precisa novamente de capacidade). Porém, se o Rebalanceamento da capacidade detectar que uma instância está sob risco de interrupção, o Amazon EC2 Auto Scaling tentará iniciar imediatamente uma nova instância. Portanto, mais instâncias poderão ser substituídas do que se você esperasse que o Amazon EC2 Auto Scaling iniciasse uma nova instância após a interrupção da instância em risco.

Embora você possa substituir mais instâncias com o rebalanceamento de capacidade habilitado, você se beneficia por ser proativo em vez de reativo. Isso lhe dá mais tempo para agir antes que suas instâncias sejam interrompidas. Com um Aviso de interrupção de instâncias spot, normalmente você só tem até dois minutos para encerrar sua instância sem problemas. Com o Rebalanceamento da capacidade iniciando uma nova instância com antecedência, os processos existentes têm maiores chances de serem concluídos na instância em risco. Além disso, você pode iniciar os procedimentos de desligamento da instância, impedir que novos trabalhos sejam agendados na instância em risco e preparar a instância recém-executada para assumir o controle da aplicação. Com a substituição proativa no reequilíbrio de capacidade, você se beneficia de uma continuidade tranquila.

O exemplo teórico a seguir demonstra os riscos e benefícios do uso do rebalanceamento de capacidade:

- 14h – Uma recomendação de rebalanceamento é recebida para a instância A. O Amazon EC2 Auto Scaling tenta imediatamente iniciar a instância de substituição B, dando a você tempo para iniciar seus procedimentos de encerramento.

- 14h30 – Uma recomendação de rebalanceamento é recebida para a instância B, que é substituída pela instância C. Isso lhe dá tempo para iniciar seus procedimentos de desligamento.
- 14h32 – Se o reequilíbrio de capacidade não estiver ativado e se um aviso de interrupção da instância spot tiver sido recebido às 14h32, para a instância A, você terá apenas dois minutos para agir. Porém, a instância A teria continuado em execução até esse momento.

Habilitação do rebalanceamento de capacidade para substituir proativamente as instâncias spot em risco

Você pode usar o Console de gerenciamento da AWS ou AWS CLI para ativar o rebalanceamento de capacidade para seu grupo de Auto Scaling. Quando o rebalanceamento de capacidade está habilitado, o Amazon EC2 Auto Scaling tenta substituir proativamente as instâncias spot no seu grupo que receberam uma recomendação de rebalanceamento de instância do EC2.

Habilitar o rebalanceamento de capacidade (console)

Você pode habilitar ou desabilitar o rebalanceamento de capacidade quando cria ou atualiza um grupo do Auto Scaling.

Para habilitar o rebalanceamento de capacidade para um novo grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Selecione Criar grupo do Auto Scaling.
3. Para Etapa 1: escolher modelo de execução ou configuração, insira um nome para o grupo do Auto Scaling, escolha um modelo de execução e escolha Próximo para prosseguir para a próxima etapa.
4. Para a Etapa 2: escolha as opções de execução da instância, em Requisitos do tipo de instância, escolha as configurações para criar um grupo misto de instâncias. Isso inclui os tipos de instância que ele pode lançar, as opções de compra de instâncias e as estratégias de alocação para instâncias sob demanda e spot. Por padrão, essas configurações não estão definidas. Para configurá-las, é necessário selecionar Override launch template (Substituir modelo de execução). Para obter mais informações sobre como criar um grupo de instâncias mistas, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

5. Em Rede, escolha as opções desejadas. Verifique se as sub-redes que você deseja utilizar se encontram em diferentes zonas de disponibilidade.
6. Na seção Estratégias de alocação, escolha uma estratégia de alocação spot. Ative ou desative o rebalanceamento de capacidade marcando ou desmarcando a caixa de seleção em Rebalanceamento de capacidade. Você só vê essa opção quando solicita que uma porcentagem do grupo do Auto Scaling seja iniciada como instâncias spot na seção Opções de compra de instância.
7. Crie o grupo do Auto Scaling.
8. (Opcional) Adicione gancho do ciclo de vida conforme necessário. Para obter mais informações, consulte Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling.

Para ativar ou desativar o reequilíbrio de capacidade para um grupo existente do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling. Um painel dividido é aberto na parte inferior da página.
3. Na guia Details (Detalhes), escolha Allocation strategies (Estratégias de alocação), Edit (Editar).
4. Na seção Estratégias de alocação, ative ou desative o rebalanceamento de capacidade marcando ou desmarcando a caixa de seleção em Rebalanceamento de capacidade.
5. Selecione Atualizar.

Habilitar o rebalanceamento de capacidade (AWS CLI)

Os exemplos a seguir mostram como usar o AWS CLI para ativar e desativar o rebalanceamento de capacidade.

Use o update-auto-scaling-groupcomando create-auto-scaling-groupou com o seguinte parâmetro:

- `--capacity-rebalance` / `--no-capacity-rebalance`: - valor booleano que indica se o rebalanceamento de capacidade está habilitado.

Antes de chamar o create-auto-scaling-groupcomando, você precisa do nome de um modelo de execução configurado para uso com um grupo de Auto Scaling. Para obter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.

> **Note**
>
> Os procedimentos a seguir mostram como usar um arquivo de configuração formatado em JSON ou YAML. Se você usar a AWS CLI versão 1, deverá especificar um arquivo de configuração formatado em JSON. Se você usar a AWS CLI versão 2, poderá especificar um arquivo de configuração formatado em YAML ou JSON.

JSON

Para criar e configurar um novo grupo do Auto Scaling

- Use o create-auto-scaling-groupcomando a seguir para criar um novo grupo de Auto Scaling e ativar o rebalanceamento de capacidade. Este comando faz referência a um arquivo JSON como único parâmetro para seu grupo de Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Se você ainda não possui um arquivo de configuração da CLI que especifique uma política de instâncias mistas, crie um.

Adicione a entrada a seguir ao objeto JSON de nível superior no arquivo de configuração.

```
{
    "CapacityRebalance": true
}
```

Veja a seguir um exemplo de arquivo `config.json`.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredCapacity": 12,
    "MinSize": 12,
    "MaxSize": 15,
    "CapacityRebalance": true,
    "MixedInstancesPolicy": {
        "InstancesDistribution": {
            "OnDemandBaseCapacity": 0,
            "OnDemandPercentageAboveBaseCapacity": 25,
            "SpotAllocationStrategy": "price-capacity-optimized"
```

```
        },
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Default"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        }
    },
    "TargetGroupARNs": "arn:aws:elasticloadbalancing:us-
west-2:123456789012:targetgroup/my-alb-target-group/943f017f100becff",
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Para criar e configurar um novo grupo do Auto Scaling

- Use o create-auto-scaling-groupcomando a seguir para criar um novo grupo de Auto Scaling e ativar o rebalanceamento de capacidade. Este comando faz referência a um arquivo YAML como único parâmetro para seu grupo do Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Adicione a linha a seguir ao arquivo de configuração formatado em YAML.

```
CapacityRebalance: true
```

Veja a seguir um exemplo de arquivo `config.yaml`.

```
---
AutoScalingGroupName: my-asg
DesiredCapacity: 12
MinSize: 12
MaxSize: 15
CapacityRebalance: true
MixedInstancesPolicy:
  InstancesDistribution:
    OnDemandBaseCapacity: 0
    OnDemandPercentageAboveBaseCapacity: 25
    SpotAllocationStrategy: price-capacity-optimized
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
TargetGroupARNs:
```

```
- arn:aws:elasticloadbalancing:us-west-2:123456789012:targetgroup/my-alb-target-
group/943f017f100becff
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Para habilitar o rebalanceamento de capacidade para um grupo do Auto Scaling existente

- Use o update-auto-scaling-groupcomando a seguir para ativar o rebalanceamento de capacidade.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --capacity-rebalance
```

Para verificar se o rebalanceamento de capacidade está habilitado para um grupo do Auto Scaling

- Use o describe-auto-scaling-groupscomando a seguir para verificar se o rebalanceamento de capacidade está ativado e para ver os detalhes.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

O seguinte é um exemplo de resposta.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            ...
            "CapacityRebalance": true
        }
    ]
}
```

Para desabilitar o rebalanceamento de capacidade

Use o update-auto-scaling-groupcomando com a `--no-capacity-rebalance` opção de desativar o rebalanceamento de capacidade.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
```

```
  --no-capacity-rebalance
```

Recursos relacionados

Para obter mais informações sobre o rebalanceamento de capacidade, consulte Gerenciar proativamente o ciclo de vida da instância spot usando o novo recurso de rebalanceamento de capacidade para Amazon EC2 Auto Scaling no blog Compute. AWS

Para obter mais informações sobre as recomendações de rebalanceamento de instâncias do EC2, consulte Recomendações de rebalanceamento de instâncias do EC2 no Guia do usuário do Amazon EC2.

Para saber mais sobre ganchos do ciclo de vida, consulte os recursos a seguir.

- Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda(usando EventBridge)
- Tutorial: Usar script de dados e metadados de instância para recuperar o estado do ciclo de vida

Limitações

- O Amazon EC2 Auto Scaling poderá substituir a instância que recebe a notificação de rebalanceamento somente se a instância não estiver protegida contra redução vertical. No entanto, a proteção escalável não impede o encerramento de uma interrupção Spot. Para obter mais informações, consulte Use a proteção de redução da escala horizontalmente da instância para controlar a terminação da instância.
- O suporte para reequilíbrio de capacidade está disponível em todas as áreas comerciais Regiões da AWS onde o Amazon EC2 Auto Scaling está disponível, exceto na região do Oriente Médio (Emirados Árabes Unidos).

Reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas

As reservas de capacidade sob demanda do Amazon EC2 permitem que você reserve a capacidade computacional em zonas de disponibilidade específicas. Para começar a usar as reservas de capacidade com o Auto Scaling, crie primeiro a reserva de capacidade ou um grupo de reserva de capacidade na zona de disponibilidade específica. Em seguida, é possível adicionar uma

preferência de reserva de capacidade ao seu grupo do Auto Scaling ao criá-lo ou ao atualizar um grupo existente.

Para criar uma reserva de capacidade, consulte Criação de uma reserva de capacidade no Guia do usuário do Amazon EC2. Para criar um grupo de reserva de capacidade, consulte Criação de um grupo de reserva de capacidade no Guia do usuário do Amazon EC2.

Preferência de reserva de capacidade

A preferência de reserva de capacidade ajuda a usar reservas de capacidade de forma eficiente, priorizando a capacidade reservada em uma reserva de capacidade antes de usar a capacidade sob demanda. É possível selecionar entre as opções de preferência de reserva de capacidade a seguir:

- Padrão: o Auto Scaling usa a preferência de reserva de capacidade do seu modelo de execução ou uma reserva de capacidade aberta.
- Nenhuma: o Auto Scaling não iniciará instâncias em uma reserva de capacidade. As instâncias serão executadas em capacidade sob demanda.
- Somente reservas de capacidade: o Auto Scaling só iniciará instâncias em uma reserva de capacidade ou grupo de reserva de capacidade. Se a capacidade não estiver disponível, as instâncias falharão ao serem iniciadas.
- Reservas de capacidade primeiro: o Auto Scaling iniciará instâncias em uma reserva de capacidade ou grupo de reserva de capacidade. Se a capacidade não estiver disponível, as instâncias serão executadas na capacidade sob demanda.

Se você selecionar "Somente reservas de capacidade" ou "Reservas de capacidade primeiro", poderá especificar um destino de reserva de capacidade.

> **Note**
>
> É necessário selecionar uma preferência de reserva de capacidade. O destino de reserva de capacidade é opcional.

Considerações sobre preferências de reserva de capacidade e modelos de execução

Considere o seguinte se você selecionar "Somente reservas de capacidade" ou "Reservas de capacidade primeiro":

- Se você selecionar "Somente reservas de capacidade" ou "Reservas de capacidade primeiro", o Auto Scaling usará a meta de reserva de capacidade especificada no grupo do Auto Scaling em vez da meta de reserva de capacidade do modelo de execução.
- Se você selecionar "Somente reservas de capacidade" ou "Reservas de capacidade primeiro" e não especificar um destino de reserva de capacidade, o Auto Scaling usará o destino de reserva de capacidade do modelo de execução ou uma reserva de capacidade aberta.

Especificação do destino de reserva de capacidade

Se você selecionar Somente reservas de capacidade ou Reservas de capacidade primeiro, as opções de destino de reserva de capacidade a seguir estarão disponíveis:

- Aberto: o Auto Scaling iniciará instâncias em qualquer reserva de capacidade aberta. Se você selecionou Somente reservas de capacidade e a capacidade não estiver disponível, as instâncias não serão iniciadas. Se você selecionou Reservas de capacidade primeiro e a capacidade não estiver disponível, as instâncias serão iniciadas na capacidade sob demanda.
- Especificar reserva de capacidade: o Auto-Scaling iniciará as instâncias na reserva de capacidade especificada. Se você selecionou Somente reservas de capacidade e a capacidade não estiver disponível, as instâncias não serão iniciadas. Se você selecionou Reservas de capacidade primeiro e a capacidade não estiver disponível, as instâncias serão iniciadas na capacidade sob demanda.
- Especificar grupo de recursos de reserva de capacidade: o Auto-Scaling iniciará as instâncias em uma reserva de capacidade aberta no grupo de recursos de reserva de capacidade especificado. Se você selecionou Somente reservas de capacidade e a capacidade não estiver disponível, as instâncias não serão iniciadas. Se você selecionou Reservas de capacidade primeiro e a capacidade não estiver disponível, as instâncias serão iniciadas na capacidade sob demanda.

Saldo da zona de disponibilidade e reservas de capacidade

O Auto Scaling prioriza o equilíbrio das zonas de disponibilidade mesmo quando você usa as preferências de reservas de capacidade. Isso significa que:

- Com `capacity-reservations-first` — O grupo Auto Scaling distribuirá as instâncias uniformemente AZs primeiro e, em seguida, usará as reservas de capacidade quando disponíveis. Se as reservas de capacidade não estiverem disponíveis em uma AZ, as instâncias serão executadas sob demanda para manter o equilíbrio da AZ.

- Com `capacity-reservations-only`: o grupo do Auto Scaling usará somente reservas de capacidade. Isso pode resultar em uma distribuição desigual de instâncias AZs com base em onde as reservas de capacidade estão disponíveis.

Exemplo

Se você tiver 10 reservas de capacidade na AZ-a, 3 na AZ-b, 1 na AZ-c e uma capacidade desejada de 9 instâncias:

- O uso de `capacity-reservations-first` para 9 instâncias resultará em 3 instâncias por AZ (mantendo o equilíbrio de AZ), com algumas instâncias potencialmente funcionando sob demanda.
- O uso de `capacity-reservations-only` para 9 instâncias resultará em uma distribuição desigual com base nas reservas de capacidade disponíveis.

Uso da preferência de reserva de capacidade com seu grupo do Auto Scaling

Para usar reservas de capacidade com seu grupo do Auto Scaling, crie primeiro uma reserva de capacidade ou um grupo recursos de reserva de capacidade. Em seguida, é possível adicionar uma preferência de reserva de capacidade ao seu grupo do Auto Scaling ao criá-lo ou ao atualizar um grupo existente.

Para criar uma reserva de capacidade, consulte Criação de uma reserva de capacidade no Guia do usuário do Amazon EC2. Para criar um grupo de reserva de capacidade, consulte Criação de um grupo de reserva de capacidade no Guia do usuário do Amazon EC2.

Use um dos métodos a seguir para usar a preferência de reserva de capacidade ao criar ou editar um grupo do Auto Scaling.

Console

Para usar a preferência de reserva de capacidade em um novo grupo (console)

1. Siga as instruções Criar um grupo do Auto Scaling usando o assistente de execução do Amazon EC2 e conclua cada etapa do procedimento, até a etapa 3.
2. Na página Configurar tamanho e escalabilidade do grupo, em Configurações adicionais de capacidade, Preferência de reserva de capacidade, selecione uma preferência de reserva

de capacidade. Para obter mais informações sobre preferências de reservas de capacidade, consulte Preferência de reserva de capacidade.

3. Continue com as etapas em Criar um grupo do Auto Scaling usando o assistente de execução do Amazon EC2.

AWS CLI

Para usar a preferência de reserva de capacidade em um novo grupo (AWS CLI)

Adicione o parâmetro `--capacity-reservation-specification` ao comando create-auto-scaling-group.

1. Especifique uma preferência de reserva de capacidade. Para obter mais informações, consulte Preferência de reserva de capacidade.
2. Especifique um destino de reserva de capacidade. Se você selecionar "Somente reservas de capacidade" ou "Reservas de capacidade primeiro" e não especificar um destino de reserva de capacidade, o Auto Scaling usará o destino de reserva de capacidade do modelo de execução ou uma reserva de capacidade aberta.

Console

Para usar a preferência de reserva de capacidade em um grupo existente (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a mesma Região da AWS na qual você criou o grupo do Auto Scaling.
3. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
4. Na guia Detalhes, em Preferência de reserva de capacidade, escolha Editar.
5. Em Configurações adicionais de capacidade, Preferência de reserva de capacidade, selecione uma preferência de reserva de capacidade. Para obter mais informações sobre preferências de reservas de capacidade, consulte Preferência de reserva de capacidade.
6. Selecione Atualizar.

AWS CLI

Para usar a preferência de reserva de capacidade em um grupo existente (AWS CLI)

Adicione o parâmetro `--capacity-reservation-specification` ao comando update-auto-scaling-group.

1. Especifique uma preferência de reserva de capacidade. Para obter mais informações, consulte Preferência de reserva de capacidade.
2. Especifique um destino de reserva de capacidade. Se você selecionar "Somente reservas de capacidade" ou "Reservas de capacidade primeiro" e não especificar um destino de reserva de capacidade, o Auto Scaling usará o destino de reserva de capacidade do modelo de execução ou uma reserva de capacidade aberta.

Reservas de capacidade interruptível com o EC2 Auto Scaling

As reservas de capacidade interruptível representam a capacidade ociosa emprestada a você pelos proprietários de reservas de capacidade sob demanda em sua organização. AWS Essas reservas são adequadas para cargas de trabalho interruptíveis porque o proprietário da capacidade pode recuperar a capacidade a qualquer momento com um aviso de interrupção de 2 minutos, após o qual as instâncias são encerradas. Para uma visão geral das reservas de capacidade interruptível e como elas funcionam, consulte Reservas de capacidade interruptível no Guia do usuário do Amazon EC2.

Para começar, você cria uma reserva interruptível alocando capacidade não utilizada de qualquer reserva de capacidade sob demanda em sua conta. Em seguida, você cria seu modelo de lançamento, especifica o ID da reserva de capacidade interruptível e o tipo de mercado apropriado. Por fim, crie seu grupo de Auto Scaling para usar o modelo de lançamento que você criou.

> **Note**
>
> - Grupos de instâncias mistas não são compatíveis com reservas de capacidade interruptível.
> - As reservas de capacidade interruptíveis aparecem em sua conta com um novo atributo `interruptible` `true` definido como.

Tópicos

- Diretrizes operacionais
- Especifique uma reserva de capacidade interruptível em seu modelo de lançamento
- Limitações
- Recursos relacionados

Diretrizes operacionais

A seguir estão as diretrizes operacionais básicas que você deve seguir ao usar reservas de capacidade interruptível com um grupo de Auto Scaling.

- As instâncias lançadas a partir de reservas de capacidade interruptível podem ser encerradas com um aviso prévio de 2 minutos quando a capacidade for recuperada. Projete seus aplicativos para lidar com desligamentos regulares.
- O Amazon EC2 envia EventBridge notificações 2 minutos antes de encerrar as instâncias. Embora o Auto Scaling não reaja automaticamente a esses eventos, você pode configurar a automação personalizada para responder aos avisos de interrupção.
- Quando as instâncias são encerradas devido à recuperação de capacidade, o verificador de integridade do Auto Scaling detectará o encerramento e iniciará automaticamente as instâncias de substituição de acordo com a configuração do seu grupo.
- Aponte para uma versão específica do modelo de lançamento em vez da `$Latest` versão `$Default` ou para garantir a consistência.

> **Note**
>
> Quando uma instância de reserva de capacidade interruptível é recuperada, as atividades de escalabilidade do seu grupo de Auto Scaling indicarão:. `taken out of service in response to an Amazon EC2 health check that indicated it had been terminated or stopped` O Auto Scaling tentará substituir a instância da mesma forma que faz com qualquer instância que falhe em uma verificação de integridade.

Especifique uma reserva de capacidade interruptível em seu modelo de lançamento

Para criar um modelo de lançamento que tenha como alvo uma reserva de capacidade interruptível específica para seu grupo de Auto Scaling, use um dos seguintes métodos:

Console

Para especificar uma reserva de capacidade interruptível em seu modelo de lançamento (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. Na barra de navegação superior, selecione Região da AWS onde você tem sua reserva de capacidade interruptível.
3. No painel de navegação, escolha Instances e, em seguida, Launch Templates.
4. Escolha Criar modelo de execução e crie o modelo de execução. Inclua o ID da AMI, o tipo de instância e qualquer outra configuração do modelo de execução, conforme necessário.
5. Amplie a seção Detalhes avançados para ver as configurações avançadas.
6. Para a opção de compra, escolha Reserva de capacidade interruptível.
7. Em Reserva de capacidade, escolha Destino por ID e, em Reserva de capacidade - Destino por ID, escolha o ID de reserva de capacidade de uma reserva de capacidade interruptível existente.
8. Quando terminar, selecione Criar modelo de execução.

Para obter ajuda na criação de um grupo de Auto Scaling com um modelo de inicialização, consulte.Criar um grupo do Auto Scaling usando um modelo de execução

AWS CLI

Para especificar uma reserva de capacidade interruptível em seu modelo de lançamento ()AWS CLI

Use o create-launch-templatecomando a seguir para criar um modelo de execução que especifica uma ID de reserva de capacidade interruptível existente. Substitua cada espaço reservado para entrada do usuário por suas próprias informações.

```
aws ec2 create-launch-template --launch-template-name my-template-for-interruptible-cr \
  --version-description AutoScalingVersion1 --region us-east-2 \
  --launch-template-data file://config.json
```

> **Tip**
>
> Se esse comando gerar um erro, verifique se você atualizou o AWS CLI localmente para a versão mais recente.

Conteúdo de `config.json`:

```
{
    "ImageId": "ami-04d5cc9b88example",
    "InstanceType": "t3.nano",
    "SecurityGroupIds": [
        "sg-903004f88example"
    ],
    "KeyName": "MyKeyPair",
    "InstanceMarketOptions": {
        "MarketType": "interruptible-capacity-reservation"
    },
    "CapacityReservationSpecification": {
        "CapacityReservationTarget": {
            "CapacityReservationId": "cr-02168da1478b509e0"
        }
    }
}
```

A seguir está um exemplo de saída:

```
{
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-068f72b724example",
        "LaunchTemplateName": "my-template-for-interruptible-cr",
        "CreateTime": "2023-10-27T15:12:44.000Z",
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "DefaultVersionNumber": 1,
        "LatestVersionNumber": 1
    }
}
```

Você pode usar o describe-launch-template-versionscomando a seguir para verificar a ID de reserva de capacidade interruptível associada ao modelo de execução:

```
aws ec2 describe-launch-template-versions --launch-template-names my-template-for-interruptible-cr \
  --region us-east-2
```

Configuração em nível de grupo do Auto Scaling

Opcionalmente, você pode especificar as configurações de reserva de capacidade no nível do grupo Auto Scaling para substituir as configurações do modelo de execução:

- Preferência de reserva de capacidade: só `capacity-reservations-only` é suportada para reservas de capacidade interruptível. A `capacity-reservations-first` preferência não é suportada.
- Especificação de destino: você pode especificar uma ID de reserva de capacidade interruptível diferente no nível do grupo para substituir a configuração do modelo de execução.

Limitações

- As políticas de instâncias mistas não são suportadas com reservas de capacidade interruptível.
- Você deve especificar o tipo de mercado correto (`interruptible-capacity-reservation`) em seu modelo de lançamento junto com o ID de reserva de capacidade.

Recursos relacionados

- Para obter informações sobre como criar e gerenciar reservas de capacidade interruptível, consulte Reservas de capacidade interruptível no Guia do usuário do Amazon EC2.
- Para obter informações gerais sobre o uso de reservas de capacidade com o Auto Scaling, consulte. Reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas
- Para obter informações sobre as verificações de saúde do grupo Auto Scaling, consulte Verificações de saúde para instâncias em um grupo do Auto Scaling.

Crie grupos de Auto Scaling a partir da linha de comando usando AWS CloudShell

Se houver suporte Regiões da AWS, você pode executar AWS CLI comandos usando AWS CloudShell um shell pré-autenticado baseado em navegador que é iniciado diretamente do. Console de gerenciamento da AWS Você pode executar AWS CLI comandos em serviços usando seu shell preferido (shell Bash ou Z). PowerShell

Você pode iniciar a AWS CloudShell partir do Console de gerenciamento da AWS usando um dos dois métodos a seguir:

- Escolha o AWS CloudShell ícone na barra de navegação do console. Ele está à direita da caixa de pesquisa.
- Use a caixa de pesquisa na barra de navegação do console para pesquisar CloudShelle, em seguida, escolha a CloudShellopção.

Quando AWS CloudShell é iniciado em uma nova janela do navegador pela primeira vez, um painel de boas-vindas é exibido e lista os principais recursos. Depois de fechar esse painel, as atualizações de status serão fornecidas enquanto o shell configura e encaminha suas credenciais do console. Quando o prompt de comando for exibido, o shell estará pronto para interação.

Para obter mais informações sobre esse serviço, consulte o Manual do usuário do AWS CloudShell.

Crie grupos de Auto Scaling com AWS CloudFormation

O Amazon EC2 Auto Scaling é AWS CloudFormation integrado com um serviço que ajuda você a modelar e AWS configurar seus recursos para que você possa gastar menos tempo criando e gerenciando seus recursos e infraestrutura. Você cria um modelo que descreve todos os AWS recursos desejados (como grupos do Auto Scaling) e CloudFormation provisiona e configura esses recursos para você.

Ao usar CloudFormation, você pode reutilizar seu modelo para configurar seus recursos do Amazon EC2 Auto Scaling de forma consistente e repetida. Descreva seus recursos uma vez e, em seguida, provisione os mesmos recursos repetidamente em várias Contas da AWS regiões.

Amazon EC2 CloudFormation Auto Scaling e modelos

Para provisionar e configurar recursos para Amazon EC2 Auto Scaling e serviços relacionados, você deve entender CloudFormation modelos. Os modelos são arquivos de texto formatados em JSON ou YAML. Esses modelos descrevem os recursos que você deseja provisionar em suas CloudFormation pilhas. Se você não estiver familiarizado com JSON ou YAML, você pode usar o CloudFormation Designer para ajudá-lo a começar a usar modelos. CloudFormation Para obter mais informações, consulte O que é CloudFormation Designer? no Guia do AWS CloudFormation usuário.

Para começar a criar seus próprios modelos de pilha para o Amazon EC2 Auto Scaling, realize as tarefas a seguir:

- Crie um modelo de lançamento usando AWS::EC2::LaunchTemplateo.
-

Para ver um passo a passo que mostra como implantar um grupo do Auto Scaling por trás de um Application Load Balancer, consulte Demonstração: criar uma aplicação escalável com balanceamento de carga no Guia do usuário do AWS CloudFormation .

Você pode encontrar exemplos úteis adicionais de modelos que criam grupos do Auto Scaling e recursos relacionados nas seções a seguir do AWS CloudFormation Guia do usuário.

- Referência de tipo de recurso do Amazon EC2 Auto Scaling
- Configure os recursos do Amazon EC2 Auto Scaling com AWS CloudFormation

Saiba mais sobre CloudFormation

Para saber mais sobre isso CloudFormation, consulte os seguintes recursos:

- AWS CloudFormation
- AWS CloudFormation Guia do usuário
- CloudFormation API Reference
- AWS CloudFormation Guia do usuário da interface de linha de comando

Obtenha recomendações de tipo de instância com AWS Compute Optimizer

AWS fornece recomendações de tipo de instância do Amazon EC2 para ajudá-lo a melhorar o desempenho, economizar dinheiro ou ambos, usando recursos desenvolvidos por. AWS Compute Optimizer Você pode usar essas recomendações para decidir se deve mudar para um novo tipo de instância no seu grupo de Auto Scaling.

Para fazer recomendações, o Compute Optimizer analisa as especificações de instância existentes e o histórico de métricas recente. Depois, os dados compilados são usados para recomendar quais tipos de instância do Amazon EC2 são mais bem otimizados para lidar com a workload de performance existente. Recomendações são retornadas com a definição de preço de instância por hora.

> **Note**
>
> Para obter recomendações do Compute Optimizer, primeiro é necessário optar pelo Compute Optimizer. Para obter mais informações, consulte Conceitos básicos do AWS Compute Optimizer no Manual do usuário do AWS Compute Optimizer .

Conteúdo

- Limitações
- Conclusões
- Exibir recomendações
- Considerações para avaliação das recomendações

Limitações

O Compute Optimizer gera recomendações para instâncias em grupos do Auto Scaling configurados para iniciar e executar os tipos de instância M, C, R, T e X. No entanto, ele não gera recomendações para tipos de instância -g alimentados por processadores AWS Graviton2 (por exemplo, C6g) e para tipos de instância -n que têm maior desempenho de largura de banda de rede (por exemplo, M5n).

Os grupos do Auto Scaling também devem ser configurados para executar um único tipo de instância (ou seja, nenhum tipo de instância mista), não devem ter uma política de escalabilidade anexada a

eles e ter os mesmos valores para a capacidade desejada, mínima e máxima (ou seja, um grupo do Auto Scaling com um número fixo de instâncias). O Compute Optimizer gera recomendações para instâncias em grupos do Auto Scaling que atendam todos esses requisitos de configuração.

Conclusões

O Compute Optimizer classifica suas descobertas para grupos do Auto Scaling da seguinte forma:

- Not optimized (Não otimizado): um grupo do Auto Scaling é considerado não otimizado quando o Compute Optimizer identifica uma recomendação que pode fornecer uma melhor performance para sua workload.
- Optimized (Otimizado): um grupo do Auto Scaling é considerado otimizado quando o Compute Optimizer determina que o grupo está provisionado corretamente para executar sua workload, com base no tipo de instância escolhido. Para recursos otimizados, o Compute Optimizer às vezes pode recomendar um tipo de instância de nova geração.
- None (Nenhum): não há recomendações para esse grupo do Auto Scaling. Isso poderá ocorrer se você tiver optado pelo Compute Optimizer há menos de 12 horas, quando o grupo do Auto Scaling estiver em execução há menos de 30 horas ou quando o grupo do Auto Scaling ou o tipo de instância não tiver suporte no Compute Optimizer. Para obter mais informações, consulte a seção Limitações.

Exibir recomendações

Depois de optar pelo Compute Optimizer, é possível visualizar as descobertas e as recomendações que ele gera para seus grupos do Auto Scaling. Caso tenho realizado a opção recentemente, as recomendações poderão não estar disponíveis durante até 12 horas.

Como visualizar as recomendações geradas para um grupo do Auto Scaling

1. Abra o console do Compute Optimizer em. https://console.aws.amazon.com/compute-optimizer/

   A página Dashboard (Painel) é aberta.
2. Escolha View recommendations for all Auto Scaling groups (Visualizar recomendações para todos os grupos de Auto Scaling).
3. Selecione seu grupo do Auto Scaling.
4. Escolha View detail (Visualizar detalhes).

A visualização muda para exibir até três recomendações de instância diferentes em uma visualização pré-configurada, com base nas configurações de tabela padrão. Ele também fornece dados CloudWatch métricos recentes (utilização média da CPU, média de entrada e média de saída de rede) para o grupo Auto Scaling.

Determine se deseja usar uma das recomendações. Decida se deseja realizar a otimização para melhorar a performance, reduzir custos ou ambos.

Para alterar o tipo de instância no grupo do Auto Scaling, atualize o modelo de execução ou o grupo do Auto Scaling para usar uma nova configuração de execução. As instâncias existentes continuam a usar a configuração anterior. Para atualizar as instâncias existentes, termine-as para que elas sejam substituídas pelo grupo do Auto Scaling, ou permita que a escalabilidade automática substitua gradualmente as instâncias mais antigas por instâncias mais novas com base em suas políticas de término.

> **Note**
>
> Com os recursos de tempo de vida máximo e atualização de instância, você também pode substituir instâncias existentes no grupo do Auto Scaling para iniciar novas instâncias que usem o modelo de execução ou a configuração de execução. Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na vida útil máxima da instância e Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling.

Considerações para avaliação das recomendações

Antes de passar para um novo tipo de instância, considere o seguinte:

- As recomendações não preveem seu uso. As recomendações são baseadas em seu histórico de uso durante os últimos 14 dias. Escolha um tipo de instância que atenda às suas necessidades de uso futuras.
- Concentre-se nas métricas gráficas para determinar se o uso real é menor do que a capacidade da instância. Você também pode visualizar dados métricos (média, pico, percentil) CloudWatch para avaliar melhor suas recomendações de instância do EC2. Por exemplo, observe como as métricas de porcentagem da CPU mudam durante o dia e se há picos que precisem ser acomodados. Para obter mais informações, consulte Visualização das métricas disponíveis no Guia CloudWatch do usuário da Amazon.

- O Compute Optimizer pode fornecer recomendações para instâncias expansíveis, que são as instâncias T3, T3a e T2. Se você ultrapassar periodicamente sua linha de base, certifique-se de continuar fazendo isso com base no v do novo CPUs tipo de instância. Para obter mais informações, consulte Créditos de CPU e desempenho de linha de base para instâncias de desempenho com capacidade de intermitência no Guia do usuário do Amazon EC2.
- Se você comprou uma Instância reservada, sua instância sob demanda poderá ser cobrada como uma Instância reservada. Antes de alterar o tipo de instância atual, avalie o impacto sobre o uso e a cobertura da Instância reservada.
- Considere conversões para instâncias da geração mais recente, sempre que possível.
- Ao migrar para uma família de instâncias diferente, verifique se o tipo de instância atual e o novo tipo de instância são compatíveis, por exemplo, em termos de virtualização, arquitetura ou tipo de rede. Para obter mais informações, consulte Compatibilidade para redimensionamento de instâncias no Guia do usuário do Amazon EC2.
- Por fim, considere a classificação de risco de performance fornecida para cada recomendação. O risco de performance indica o esforço necessário para validar se o tipo de instância recomendado atende aos requisitos de performance da sua workload. Também recomendamos testes rigorosos de carga e performance antes e depois de fazer quaisquer alterações.

Recursos adicionais do

Além dos tópicos desta página, consulte os seguintes recursos:

- Tipos de instância do Amazon EC2
- AWS Compute Optimizer Guia do usuário

Use o Elastic Load Balancing para distribuir automaticamente o tráfego de entrada de aplicações no seu grupo do Auto Scaling.

O Elastic Load Balancing distribui automaticamente o tráfego de entrada de aplicações entre todas as instâncias do EC2 em execução. O Elastic Load Balancing ajuda a gerenciar solicitações de entrada roteando o tráfego de forma otimizada para que nenhuma instância seja sobrecarregada. Para usar o Elastic Load Balancing com seu grupo do Auto Scaling, anexe o balanceador de carga ao seu grupo do Auto Scaling. Isso registra o grupo com o balanceador de carga, o qual atua como um ponto único de contato para todo o tráfego da Web de entrada para seu grupo do Auto Scaling.

Quando você usa o Elastic Load Balancing com seu grupo do Auto Scaling, não é necessário registrar suas instâncias do EC2 no balanceador de carga. As instâncias iniciadas pelo grupo do Auto Scaling serão automaticamente registradas no balanceador de carga. Da mesma forma, as instâncias que são terminadas pelo grupo do Auto Scaling terão o registro cancelado automaticamente no balanceador de carga.

Depois de anexar um load balancer ao grupo do Auto Scaling, você poderá configurar o grupo do Auto Scaling para usar métricas do Elastic Load Balancing (como a contagem de solicitações do Application Load Balancer por destino) para dimensionar o número de instâncias no grupo conforme a demanda flutua.

Opcionalmente, você pode adicionar verificações de integridade do Elastic Load Balancing ao seu grupo do Auto Scaling para que o Amazon EC2 Auto Scaling possa identificar e substituir instâncias não íntegras com base nessas verificações de integridade adicionais. Caso contrário, você pode criar um CloudWatch alarme que o notifique se a contagem de hosts saudáveis do grupo-alvo for menor do que a permitida.

Conteúdo

- Tipos de Elastic Load Balancing
- Para anexar um balanceador de carga do Elastic Load Balancing a uma camada
- Este tópico descreve como anexar um balanceador de carga Elastic Load Balancing ao grupo do Auto Scaling.
- Configurar um Application Load Balancer ou Network Load Balancer pelo console
- Verificar o status do anexo de seu balanceador de carga
- Adicione uma Zona de disponibilidade
- Remover uma Zona de disponibilidade
- Desanexar um grupo de destino ou um Classic Load Balancer do seu grupo do Auto Scaling
- Exemplos para trabalhar com o Elastic Load Balancing usando o AWS CLI

Tipos de Elastic Load Balancing

O Elastic Load Balancing oferece quatro tipos de balanceadores de carga que podem ser usados com seu grupo do Auto Scaling: balanceadores de carga de aplicação, balanceadores de carga de rede, balanceadores de carga de gateway e balanceadores de carga clássicos

Há uma diferença fundamental em como os tipos de balanceadores de carga são configurados. Com os balanceadores de carga de aplicação, balanceadores de carga de rede e balanceadores de carga de gateway, as instâncias são registradas como destinos com um grupo de destino, e o tráfego deve ser roteado para o grupo de destino. Com balanceadores de carga clássicos, as instâncias são registradas diretamente no balanceador de carga.

Application Load Balancer

Roteia e faz balanceamento de carga na camada da aplicação (HTTP/HTTPS) e é compatível com roteamento baseado em caminho. Um Application Load Balancer pode rotear solicitações para portas em um ou mais destinos registrados, como instâncias do EC2, na sua nuvem privada virtual (VPC).

Network Load Balancer

Roteia e promove o balanceamento de carga na camada de transporte (camada 4 do TCP/UDP) com base nas informações de endereço extraídas do cabeçalho da camada 4. Os balanceadores de carga de rede podem lidar com picos de tráfego, reter o IP de origem do cliente e usar um IP fixo para a vida útil do balanceador de carga.

Balanceador de carga de gateway

Distribui o tráfego para uma frota de instâncias de dispositivos. Fornece escalabilidade, disponibilidade e simplicidade para dispositivos virtuais de terceiros, como firewalls, sistemas de detecção e prevenção de intrusões e outros dispositivos. Os balanceadores de carga de gateway funcionam com dispositivos virtuais compatíveis com o protocolo GENEVE. Integração técnica adicional é necessária, portanto, certifique-se de consultar o manual do usuário antes de escolher um balanceador de carga de gateway.

Classic Load Balancer

Rotas e balanceamentos de carga na camada de transporte (TCP/SSL), or at the application layer (HTTP/HTTPS).

Para obter uma compreensão mais profunda dos diferentes tipos de balanceadores de carga disponíveis, consulte os recursos a seguir:

- O que é Elastic Load Balancing?
- O que é um Application Load Balancer?
- O que é um Network Load Balancer?

- O que é um balanceador de carga de gateway?
- O que é um Classic Load Balancer?

Para anexar um balanceador de carga do Elastic Load Balancing a uma camada

É necessário cumprir os pré-requisitos para anexar um balanceador de carga a seu grupo do Auto Scaling.

- Isso inclui criar o balanceador de carga e o grupo de destino que será usado para encaminhar o tráfego ao grupo do Auto Scaling.

  Há duas maneiras de criar o balanceador de carga e o grupo de destino:

  - Como usar o Elastic Load Balancing: Siga os procedimentos na documentação do Elastic Load Balancing para criar e configurar o balanceador de carga e o grupo de destino antes de criar o grupo do Auto Scaling. Ignore a etapa para registrar suas instâncias do Amazon EC2. O Amazon EC2 Auto Scaling se encarrega automaticamente de registrar (e cancelar o registro) as instâncias quando você anexa um grupo de destino ao seu grupo do Auto Scaling. Para obter mais informações, consulte Conceitos básicos do Elastic Load Balancing no Manual do usuário do Elastic Load Balancing.
  - Como usar o Amazon EC2 Auto Scaling: Crie, configure e anexe o balanceador de carga e o grupo de destino com uma configuração básica no console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte Configurar um Application Load Balancer ou Network Load Balancer pelo console.
- Antes de criar um balanceador de carga, saiba o tipo que você precisa.. Para obter mais informações, consulte Tipos de Elastic Load Balancing.
- O balanceador de carga e seu grupo-alvo devem estar na mesma Conta da AWS VPC e região do seu grupo de Auto Scaling.
- O grupo de destino deve especificar um tipo de destino `instance`. Não é possível especificar um tipo de destino `ip` ao usar um grupo do Auto Scaling.
- Se seu modelo de execução não contém um grupo de segurança que permita que o tráfego do balanceador de carga alcance seu grupo do Auto Scaling, é necessário atualizar o modelo de execução. As regras recomendadas dependem do tipo de balanceador de carga e dos tipos de backends por ele usados. Por exemplo, para rotear o tráfego para servidores Web, permita o acesso HTTP de entrada na porta 80 a partir do balanceador de carga. As instâncias existentes

não são atualizadas com as novas configurações quando o modelo de execução é modificado. Para atualizar as instâncias existentes, você pode iniciar uma atualização para substituir a instância. Para obter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling.

- Os grupos de segurança que você especifica no modelo de execução deve permitir o acesso do balanceador de carga na porta correta para que o Elastic Load Balancing execute suas verificações de integridade.
- Ao implantar dispositivos virtuais atrás de um Load Balancer de gateway, a imagem de máquina da Amazon (AMI) deve especificar o ID de uma AMI compatível com o protocolo GENEVE para permitir que o grupo do Auto Scaling troque tráfego com um balanceador de carga Gateway. Além disso, os grupos de segurança especificados no modelo de execução ou na configuração de execução devem permitir o tráfego UDP na porta 6081.

> **Tip**
>
> Se você tiver scripts de bootstrap que levam um tempo para serem concluídos, você pode adicionar opcionalmente um gancho do ciclo de vida de execução ao seu grupo do Auto Scaling para atrasar o registro das instâncias atrás do balanceador de carga antes que seus scripts de bootstrap sejam concluídos com êxito e as aplicações nas instâncias estejam prontas para aceitar o tráfego. Você não pode adicionar um gancho do ciclo de vida ao criar inicialmente um grupo do Auto Scaling no console do Amazon EC2 Auto Scaling. Você pode adicionar um gancho do ciclo de vida após a criação do grupo. Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

Configurar verificações de integridade para destinos

É possível configurar verificações de integridade para os destinos registrados com um balanceador de carga do Elastic Load Balancing para garantir que eles sejam capazes de lidar com o tráfego adequadamente. As etapas específicas variam de acordo com o tipo de balanceador de carga que você está usando. Para saber mais, consulte os seguintes recursos:

- Application Load Balancer: Consulte Verificações de integridade para seus grupos de destino no Manual do usuário de Application Load Balancers.
- Network Load Balancer: Consulte Verificações de integridade para seus grupos de destino no Manual do usuário de Network Load Balancers.

- Gateway Load Balancer: Consulte Verificações de integridade para seus grupos de destino no Manual do usuário de Gateway Load Balancers.
- Classic Load Balancer: Consulte Configurar verificações de integridade para seu Classic Load Balancer no Manual do usuário de Classic Load Balancers.

Por padrão, o Amazon EC2 Auto Scaling não considerará uma instância não íntegra e a substituirá se ela apresentar falha nas verificações de integridade do Elastic Load Balancing. As verificações de integridade padrão para um grupo do Auto Scaling são somente verificações de integridade do EC2. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Para permitir que o Amazon EC2 Auto Scaling substitua as instâncias que são indicadas como não íntegras pelo Elastic Load Balancing, você pode configurar seu grupo do Auto Scaling para usar as verificações de integridade do Elastic Load Balancing. Ao fazer isso, o Amazon EC2 Auto Scaling considera a instância não íntegra se ela falhar nas verificações de integridade do EC2 ou nas verificações de integridade do Elastic Load Balancing. Se você anexar vários grupos de destino do balanceador de carga ou balanceadores de carga clássicos ao grupo, todos eles deverão informar que a instância é íntegra para que ela seja considerada íntegra. Se um deles relatar uma instância como não íntegra, o grupo do Auto Scaling substituirá a instância, mesmo que outros a relatem como íntegra.

Para obter mais informações sobre como habilitar as verificações de integridade do Elastic Load Balancing para seu grupo do Auto Scaling, consulte Este tópico descreve como anexar um balanceador de carga Elastic Load Balancing ao grupo do Auto Scaling..

> **Note**
>
> Para garantir que essas verificações de integridade comecem o mais rápido possível, não defina um período de carência da verificação de integridade do grupo muito alto, mas alto o suficiente para que suas verificações de integridade do Elastic Load Balancing consigam determinar se um destino está disponível para lidar com solicitações. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

Este tópico descreve como anexar um balanceador de carga Elastic Load Balancing ao grupo do Auto Scaling.

Este tópico descreve como anexar um balanceador de carga Elastic Load Balancing ao grupo do Auto Scaling. Também descreve como ativar as verificações de integridade do VPC Lattice para permitir que o Amazon EC2 Auto Scaling substitua instâncias que o VPC Lattice relata como não íntegras.

Por padrão, o Amazon EC2 Auto Scaling substitui somente instâncias que não sejam íntegras ou inacessíveis com base nas verificações de integridade do Amazon EC2. Se você ativar as verificações de integridade do VPC Lattice, o Amazon EC2 Auto Scaling poderá substituir uma instância em execução se algum dos grupos de destino do VPC Lattice que você anexar ao grupo do Auto Scaling relatar que ela não é íntegra.

Para acessar um tutorial sobre como anexar um Application Load Balancer ao seu grupo do Auto Scaling, consulte Tutorial: Configurar uma aplicação escalonada e com balanceamento de carga.

> **Important**
>
> Antes de continuar, preencha todos os pré-requisitos na seção anterior.

Conteúdo

- Anexar seu grupo-alvo ou Classic Load Balancer
- Desvincular seu grupo-alvo ou Classic Load Balancer

Anexar seu grupo-alvo ou Classic Load Balancer

Ao criar ou atualizar um grupo do Auto Scaling, você pode anexar um ou mais grupos de destino ou Classic Load Balancers. Ao anexar um Application Load Balancer, Network Load Balancer ou balanceador de carga de gateway, você anexa um grupo de destino.

Siga as etapas nesta seção para usar o console para:

- Anexar um grupo de destino ou um Classic Load Balancer a um grupo do Auto Scaling
- Ativar as verificações de integridade do Elastic Load Balancing

Para anexar um balanceador de carga existente enquanto cria um novo grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a Região da AWS em que sua pilha foi criada.
3. Selecione Criar grupo do Auto Scaling.
4. Nas etapas 1 e 2, escolha as opções conforme desejado e prossiga para Etapa 3: Configurar opções avançadas.
5. Em Load balancing (Balanceamento de carga), escolha Attach to an existing load balancer (Anexar a um balanceador de carga existente).
6. Em Attach to an existing load balancer (Anexar a um balanceador de carga existente), siga um destes procedimentos:

   a. Para balanceadores de carga de aplicação, balanceadores de carga de rede e balanceadores de carga de gateway, especifique a propriedade:

      Escolha Choose from your load balancer target groups (Escolher entre os grupos de destino do balanceador de carga) e, em seguida, escolha um grupo de destino no campo Existing load balancer target groups (Grupos de destino do balanceador de carga existente).

   b. Para balanceadores de carga clássicos:

      Escolha Choose from Classic Load Balancers (Escolher entre balanceadores de carga clássicos) e, em seguida, escolha seu balanceador de carga no campo Classic Load Balancers (Balanceadores de carga clássicos).
7. (Opcional) Para verificações de integridade e tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Elastic Load Balancing.
8. Opcional em Tempo de carência da verificação de integridade, insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.
9. Prossiga para criar o grupo do Auto Scaling. Suas instâncias serão registradas automaticamente no balanceador de carga após a criação do grupo do Auto Scaling.

Para anexar um balanceador de carga existente ao seu grupo do Auto Scaling após ter sido criado

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Integrações, escolha Balanceamento de carga, Editar.
4. Em Load balancing (Balanceamento de carga), siga um destes procedimentos:

   a. Para Application, Network or Gateway Load Balancer target groups (Grupos de balanceadores de carga de aplicação, rede ou gateway), marque sua caixa de seleção e escolha um grupo de destino.

   b. Para Classic Load Balancers (Balanceadores de carga clássicos), marque sua caixa de seleção e escolha seu balanceador de carga.
5. Selecione Atualizar.

Ao terminar de anexar o grupo de destino, você pode, opcionalmente, ativar as verificações de integridade que o usam.

Para adicionar verificações de integridade do Elastic Load Balancing

1. Na guia Detalhes, escolha Verificações de integridade, Editar.
2. Para Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Elastic Load Balancing.
3. Em Período de carência da verificação de integridade, insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.
4. Selecione Atualizar.

> **Note**
>
> Você pode usar a AWS CLI para monitorar o status do balanceador de carga enquanto ele está sendo conectado. Quando o Amazon EC2 Auto Scaling registra com êxito as instâncias e pelo menos uma instância registrada passa nas verificações de integridade, você recebe o status de `InService`. Para obter mais informações, consulte Verificar o status do anexo de seu balanceador de carga.

Desvincular seu grupo-alvo ou Classic Load Balancer

Quando o balanceador de carga não for mais necessário, use o procedimento a seguir para desvinculá-lo do grupo do Auto Scaling.

Para desvincular um balanceador de carga de um grupo

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado de um grupo existente.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Detalhes, escolha Balanceamento de carga, Editar.
4. Em Load balancing (Balanceamento de carga), siga um destes procedimentos:

   a. Em Application, Network or Gateway Load Balancer target groups (Grupos de balanceadores de carga de aplicação, rede ou gateway), escolha o ícone de exclusão (X) próximo ao grupo de destino.

   b. Em Classic Load Balancers (Balanceadores de carga clássicos), escolha o ícone de exclusão (X) próximo ao balanceador de carga.
5. Selecione Atualizar.

Ao terminar de desanexar o grupo de destino, você pode desativar as verificações de integridade do VPC Lattice.

Para adicionar verificações de integridade do Elastic Load Balancing

1. Na guia Detalhes, escolha Verificações de integridade, Editar.

2. Para Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Elastic Load Balancing.
3. Selecione Atualizar.

Configurar um Application Load Balancer ou Network Load Balancer pelo console

Use o procedimento a seguir para criar e anexar um Application Load Balancer ou um Network Load Balancer enquanto cria o grupo do Auto Scaling.

Para criar anexar um novo balanceador de carga enquanto cria um novo grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Selecione Criar grupo do Auto Scaling.
3. Nas etapas 1 e 2, escolha as opções conforme desejado e prossiga para Etapa 3: Configurar opções avançadas.
4. Em Load balancing (Balanceamento de carga), escolha Attach to an new load balancer (Anexar a um novo balanceador de carga).
    a. Em Attach to a new load balancer (Anexar a um novo balanceador de carga), para Load balancer type (Tipo de balanceador de carga), escolha se deseja criar um Application Load Balancer ou Network Load Balancer
    b. Em Load balancer name (Nome do balanceador de carga), insira um nome para o balanceador de carga ou mantenha o nome padrão.
    c. Em Load balancer scheme (Esquema do balanceador de carga), escolha se deseja criar um balanceador de carga voltado para a Internet pública ou mantenha o padrão para criar um balanceador de carga interno.
    d. Em Availability Zones and subnets (Zonas de disponibilidade e sub-redes), selecione a sub-rede pública para cada zona de disponibilidade em que você optou por iniciar suas instâncias do EC2. (Estes são pré-preenchidos pela etapa 2.).
    e. Em Listeners e routing (Listeners e roteamento), atualize o número da porta do listener (se necessário) e, em Default routing (Roteamento padrão), escolha Create a target group (Criar um grupo de destino). Alternativamente, você pode escolher um grupo de destino existente na lista suspensa.

f. Se você escolheu Create a target group (Criar um grupo de destino) na última etapa, em New target group name (Nome do novo grupo de destino), insira um nome para o grupo de destino ou mantenha o nome padrão.

g. Para adicionar etiquetas ao balanceador de carga, escolha Add tag (Adicionar etiqueta) e forneça uma chave de etiqueta e um valor para cada etiqueta.

5. (Opcional) Para verificações de integridade e tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Elastic Load Balancing.

6. Opcional em Tempo de carência da verificação de integridade, insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

7. Prossiga para criar o grupo do Auto Scaling. Suas instâncias serão registradas automaticamente no balanceador de carga após a criação do grupo do Auto Scaling.

> **Note**
>
> Depois de criar o grupo do Auto Scaling, você poderá usar o console do Elastic Load Balancing para criar listeners adicionais. Isso será útil se você precisar criar um listener com um protocolo seguro, como HTTPS ou um ouvinte UDP. Você pode adicionar mais listeners aos balanceadores de carga existentes, desde que use portas distintas.

Verificar o status do anexo de seu balanceador de carga

Após você associar um balanceador de carga, ele entra no estado `Adding` ao registrar as instâncias no grupo. Quando todas as instâncias do grupo são registradas, ele entra no estado`Added`. Depois que pelo menos uma instância registrada passa nas verificações de integridade, ele entra no estado `InService`. Após o balanceador de carga entrar no estado `InService`, o Amazon EC2 Auto Scaling pode encerrar e substituir todas as instâncias relatadas como não íntegras. Se nenhuma instância registrada passar nas verificações de integridade (por exemplo, devido a um erro na configuração da verificação de integridade), o balanceador de carga não entrará no estado `InService`. O Amazon EC2 Auto Scaling não termina e substitui as instâncias.

Quando você desanexa um balanceador de carga, ele entra no estado `Removing` ao cancelar o registro das instâncias do grupo. As instâncias permanecem em execução depois que seus

registros são cancelados. Por padrão, a descarga da conexão (atraso de cancelamento de registro) é habilitada para Application Load Balancers, Network Load Balancers e Gateway Load Balancers. Se a descarga de conexão estiver habilitada, o Elastic Load Balancing aguardará que as solicitações em andamento sejam concluídas ou que o limite de tempo máximo expire (o que ocorrer primeiro) antes de cancelar o registro das instâncias.

Você pode verificar o status do anexo usando o AWS Command Line Interface (AWS CLI) ou AWS SDKs. Você não pode verificar o status do anexo no console.

Para usar o AWS CLI para verificar o status do anexo

O describe-traffic-sourcescomando a seguir retorna o status do anexo de todas as fontes de tráfego do grupo de Auto Scaling especificado.

```
aws autoscaling describe-traffic-sources --auto-scaling-group-name my-asg
```

O exemplo retorna o ARN do grupo-alvo do Elastic Load Balancing que está vinculado ao grupo do Auto Scaling, junto com o status do anexo do grupo-alvo no elementoState.

```
{
    "TrafficSources": [
        {
            "Identifier": "arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/1234567890123456",
            "State": "InService",
            "Type": "elbv2"
        }
    ]
}
```

Adicione uma Zona de disponibilidade

Para se beneficiar da segurança e da confiabilidade da redundância geográfica, distribua seu grupo do Auto Scaling em várias zonas de disponibilidade dentro de uma região e anexe um balanceador de carga para distribuir o tráfego de entrada entre essas zonas de disponibilidade.

Quando uma zona de disponibilidade se torna não íntegra ou indisponível, o Amazon EC2 Auto Scaling inicia novas instâncias em uma zona de disponibilidade não afetada. Quando a zona de disponibilidade não íntegra retornar para um estado íntegro, o Amazon EC2 Auto Scaling redistribuirá

automaticamente as instâncias da aplicação uniformemente entre todas as zonas de disponibilidade designadas. O Amazon EC2 Auto Scaling faz isso tentando iniciar novas instâncias na zona de disponibilidade com o menor número de instâncias. No entanto, se a tentativa falhar, o Amazon EC2 Auto Scaling tentará iniciá-las em outras zonas de disponibilidade até obter êxito.

O Elastic Load Balancing cria um nó de balanceador de carga para cada zona de disponibilidade que você habilita para o balanceador de carga. Se você habilitar o balanceamento de carga entre zonas, cada nó do balanceador de carga distribuirá o tráfego uniformemente entre as instâncias registradas em todas as zonas de disponibilidade habilitadas. Se o balanceamento de carga entre zonas estiver desabilitado, cada nó do balanceador de carga distribuirá solicitações uniformemente às instâncias registradas somente em sua zona de disponibilidade.

É necessário especificar pelo menos uma zona de disponibilidade ao criar seu grupo do Auto Scaling. Posteriormente, você poderá expandir a disponibilidade da sua aplicação adicionando uma zona de disponibilidade ao seu grupo do Auto Scaling e habilitando essa zona de disponibilidade para seu balanceador de carga (se o balanceador de carga oferecer suporte a ela).

Limitações

Para atualizar quais zonas de disponibilidade estão habilitadas para seu balanceador de carga, é necessário estar ciente das seguintes limitações:

- Ao habilitar uma zona de disponibilidade para seu balanceador de carga, você especifica uma sub-rede nessa zona de disponibilidade. Observe que é possível habilitar no máximo uma sub-rede por zona de disponibilidade para seu balanceador de carga.
- Para balanceadores de carga voltados para a Internet, as sub-redes especificadas para o balanceador de carga devem ter pelo menos oito endereços IP disponíveis.
- Para balanceadores de carga de aplicação, é necessário habilitar pelo menos duas zonas de disponibilidade.
- Para balanceadores de carga de rede, você não pode desabilitar as zonas de disponibilidade habilitadas, mas pode habilitar zonas adicionais.
- Para balanceadores de carga de rede, você não pode desabilitar as zonas de disponibilidade habilitadas, mas pode habilitar zonas adicionais.

Use o procedimento a seguir para expandir seu grupo do Auto Scaling e balanceador de carga para uma sub-rede em uma zona de disponibilidade adicional.

Para adicionar uma zona de disponibilidade

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado de um grupo existente.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Detalhes, escolha Rede, Editar.
4. Em Subnets (Sub-redes), escolha a sub-rede correspondente à zona de disponibilidade que deseja adicionar ao grupo do Auto Scaling.
5. Selecione Atualizar.
6. Para atualizar as zonas de disponibilidade do seu balanceador de carga para que ele compartilhe as mesmas zonas do seu grupo do Auto Scaling, execute as seguintes etapas:

   a. No painel de navegação, em Load Balancing (Balanceamento de carga), escolha Load balancers (Balanceadores de carga).

   b. Escolha seu balanceador de carga.

   c. Execute um destes procedimentos:

      - Para balanceadores de carga de aplicação e balanceadores de carga de rede:

        1. Na guia Description (Descrição), em Availability Zones (Zonas de disponibilidade), escolha Edit subnets (Editar sub-redes).
        2. Na página Edit subnets (Editar sub-redes), para Availability Zones (Zonas de disponibilidade), marque a caixa de seleção para a zona de disponibilidade a ser adicionada. Se houver somente uma sub-rede para essa zona de disponibilidade, ela estará selecionada. Se houver mais de uma sub-rede para essa zona de disponibilidade, selecione uma das opções disponíveis.

      - Para balanceadores de carga clássicos em uma VPC:

        1. Na guia Instâncias, selecione Editar zonas de disponibilidade.
        2. Na página Add and Remove Subnets (Adicionar e remover sub-redes), em Available subnets (Sub-redes disponíveis), selecione a sub-rede usando o ícone de adicionar (+). A sub-rede é movida sob Sub-redes selecionadas.

   d. Escolha Salvar.

Recursos relacionados

O Amazon EC2 Auto Scaling rebalanceia seu grupo quando você altera as zonas de disponibilidade. Isso significa substituir e redistribuir algumas instâncias. Para obter mais informações, consulte Exemplo: distribuir instâncias entre zonas de disponibilidade.

Se você registrou destinos em zonas de disponibilidade que não estão habilitados para o balanceador de carga, o balanceador de carga não roteará o tráfego para eles. Para obter mais informações, consulte Como o Elastic Load Balancing funciona no Manual do usuário do Elastic Load Balancing.

Remover uma Zona de disponibilidade

Para remover uma zona de disponibilidade do grupo do Auto Scaling e do balanceador de carga, use o procedimento a seguir.

Para remover uma zona de disponibilidade

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado de um grupo existente.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Detalhes, escolha Rede, Editar.
4. Em Subnets (Sub-redes), escolha o ícone de exclusão (X) para a sub-rede correspondente à zona de disponibilidade que deseja remover do grupo do Auto Scaling. Se houver mais de uma sub-rede para essa zona, escolha o ícone de exclusão (X) para cada uma delas.
5. Selecione Atualizar.
6. Para atualizar as zonas de disponibilidade do seu balanceador de carga para que ele compartilhe as mesmas zonas do seu grupo do Auto Scaling, execute as seguintes etapas:

   a. No painel de navegação, em Load Balancing (Balanceamento de carga), escolha Load balancers (Balanceadores de carga).

   b. Escolha seu balanceador de carga.

   c. Execute um destes procedimentos:

      - e para Application Load Balances

1. Na guia Description (Descrição), em Availability Zones (Zonas de disponibilidade), escolha Edit subnets (Editar sub-redes).
2. Na página Edit subnets (Editar sub-redes), para Availability Zones (Zonas de disponibilidade), desmarque a caixa de seleção para remover a sub-rede da zona de disponibilidade selecionada.

- Para balanceadores de carga clássicos em uma VPC:

1. Na guia Instâncias, selecione Editar zonas de disponibilidade.
2. Na página Add and Remove Subnets (Adicionar e remover sub-redes), em Available subnets (Sub-redes disponíveis), remova a sub-rede usando o ícone de exclusão (-). A sub-rede é movida para Available subnets (Sub-redes disponíveis).

d. Escolha Salvar.

Desanexar um grupo de destino ou um Classic Load Balancer do seu grupo do Auto Scaling

Quando o balanceador de carga não for mais necessário, use o procedimento a seguir para desvinculá-lo do grupo do Auto Scaling.

Para desvincular um balanceador de carga de um grupo

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado de um grupo existente.

   Um painel dividido é aberto na parte inferior da página Auto Scaling groups (Grupos do Auto Scaling).
3. Na guia Detalhes, escolha Balanceamento de carga, Editar.
4. Em Load balancing (Balanceamento de carga), siga um destes procedimentos:

   a. Em Application, Network or Gateway Load Balancer target groups (Grupos de balanceadores de carga de aplicação, rede ou gateway), escolha o ícone de exclusão (X) próximo ao grupo de destino.

   b. Em Classic Load Balancers (Balanceadores de carga clássicos), escolha o ícone de exclusão (X) próximo ao balanceador de carga.

5. Selecione Atualizar.

Ao terminar de desanexar o grupo de destino, você pode desativar as verificações de integridade do VPC Lattice.

Para adicionar verificações de integridade do Elastic Load Balancing

1. Na guia Detalhes, escolha Verificações de integridade, Editar.
2. Para Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Elastic Load Balancing.
3. Selecione Atualizar.

Exemplos para trabalhar com o Elastic Load Balancing usando o AWS CLI

Use o AWS Command Line Interface (AWS CLI) para anexar, separar e descrever balanceadores de carga e grupos-alvo, adicionar e remover verificações de saúde do Elastic Load Balancing e alterar quais zonas de disponibilidade estão habilitadas.

Este tópico mostra exemplos de AWS CLI comandos que executam tarefas comuns para o Amazon EC2 Auto Scaling.

> **Important**
>
> Para obter exemplos de comandos adicionais, consulte `aws elbv2` e `aws elb` na Referência de comandos AWS CLI .

Conteúdo

- Anexar seu grupo-alvo ou Classic Load Balancer
- Descreva seus grupos de destino ou Classic Load Balancers
- Adicionar verificações de integridade do Elastic Load Balancing
- Alterar suas zonas de disponibilidade
- Desvincular seu grupo-alvo ou Classic Load Balancer
- Remover as verificações de integridade do Elastic Load Balancing
- Comandos legados

Anexar seu grupo-alvo ou Classic Load Balancer

Use o create-auto-scaling-groupcomando a seguir para criar um grupo de Auto Scaling e anexar simultaneamente um grupo-alvo especificando seu Amazon Resource Name (ARN). O grupo de destino pode ser associado a um Application Load Balancer, um Network Load Balancer ou um balanceador de carga do Gateway.

Substitua os valores de amostra por,`--auto-scaling-group-name`,`--vpc-zone-identifier` `--min-size` e `--max-size`. Para a opção,`--launch-template` substitua *`my-launch-template`* e *`1`* pelo nome e versão de um modelo de execução para seu grupo do Auto Scaling. Para a opção,`--traffic-sources` substitua o ARN de amostra pelo ARN de um grupo de destino para um Application Load Balancer, Network Load Balancer ou balanceador de carga de gateway.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --min-size 1 --max-size 5 \
  --traffic-sources "Identifier=arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE1"
```

Use o attach-traffic-sourcescomando para anexar outros grupos-alvo ao grupo Auto Scaling depois que ele for criado.

O comando a seguir adiciona outro grupo-alvo ao mesmo grupo.

```
aws autoscaling attach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE2"
```

Como alternativa, para anexar um Classic Load Balancer ao seu grupo, especifique as opções `--traffic-sources` e `--type` ao usar create-auto-scaling-group ou,attach-traffic-sources como no exemplo a seguir. Substitua *`my-classic-load-balancer`* pelo nome de um Classic Load Balancer. Para a opção,`--type` especifique um valor de **`elb`**.

```
--traffic-sources "Identifier=my-classic-load-balancer" --type elb
```

Descreva seus grupos de destino ou Classic Load Balancers

Para descrever os balanceadores de carga ou os grupos-alvo vinculados ao seu grupo de Auto Scaling, use o comando a describe-traffic-sourcesseguir. Substitua *`my-asg`* pelo nome do seu grupo.

```
aws autoscaling describe-traffic-sources --auto-scaling-group-name my-asg
```

O exemplo retorna o ARN dos grupos de destino do Elastic Load Balancing que você anexou ao grupo do Auto Scaling.

```
{
    "TrafficSources": [
        {
            "Identifier": "arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE1",
            "State": "InService",
            "Type": "elbv2"
        },
        {
            "Identifier": "arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE2",
            "State": "InService",
            "Type": "elbv2"
        }
    ]
}
```

Para ver uma explicação do campo `State` na saída, consulte Verificar o status do anexo de seu balanceador de carga.

Adicionar verificações de integridade do Elastic Load Balancing

Para adicionar as verificações de saúde do Elastic Load Balancing às verificações de saúde que seu grupo de Auto Scaling realiza nas instâncias, use o comando a update-auto-scaling-groupseguir e **ELB** especifique como o valor da opção. `--health-check-type` Substitua *my-asg* pelo nome do seu grupo.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "ELB"
```

As novas instâncias geralmente precisam de tempo para um breve aquecimento antes de passarem por uma verificação de integridade. Se o período de carência não proporcionar tempo de aquecimento suficiente, as instâncias poderão não parecer prontas para servir tráfego. O Amazon EC2 Auto Scaling pode considerar essas instâncias não íntegras e substituí-las.

Para atualizar o período de carência da verificação de integridade, use a opção `--health-check-grace-period` ao usar,update-auto-scaling-group como no exemplo a seguir. *300*Substitua pelo número de segundos para manter as novas instâncias em serviço antes de encerrá-las, caso não estejam íntegras.

```
--health-check-grace-period 300
```

Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Alterar suas zonas de disponibilidade

Alterar suas zonas de disponibilidade apresenta algumas limitações das quais é necessário estar ciente. Para obter mais informações, consulte Adicione uma Zona de disponibilidade.

Para alterar as zonas de disponibilidade de um Application Load Balancer ou Network Load Balancer

1. Antes de alterar as zonas de disponibilidade do balanceador de carga, é uma boa ideia primeiro atualizar as zonas de disponibilidade do grupo do Auto Scaling para verificar se há disponibilidade para seus tipos de instância nas zonas especificadas.

   Para atualizar as zonas de disponibilidade do seu grupo de Auto Scaling, use o comando a seguir update-auto-scaling-group. Substitua a sub-rede IDs de amostra pela IDs das sub-redes nas zonas de disponibilidade a serem ativadas. As sub-redes especificadas substituem as sub-redes habilitadas anteriormente. Substitua *my-asg* pelo nome do seu grupo.

   ```
   aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
     --vpc-zone-identifier "subnet-41767929,subnet-cb663da2,subnet-8360a9e7"
   ```

2. Use o describe-auto-scaling-groupscomando a seguir para verificar se as instâncias nas novas sub-redes foram iniciadas. Se as instâncias tiverem sido iniciadas, você verá uma lista das instâncias e seus status. Substitua *my-asg* pelo nome do seu grupo.

   ```
   aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
   ```

3. Use o comando set-subnets a seguir para especificar as sub-redes do seu balanceador de carga. Substitua a sub-rede IDs de amostra pela IDs das sub-redes nas zonas de disponibilidade a serem ativadas. É possível especificar somente uma sub-rede por Zona de disponibilidade. As sub-redes especificadas substituem as sub-redes habilitadas anteriormente. Substitua *my-lb-arn* pelo ARN do seu balanceador de carga.

```
aws elbv2 set-subnets --load-balancer-arn my-lb-arn \
  --subnets subnet-41767929 subnet-cb663da2 subnet-8360a9e7
```

Para alterar as zonas de disponibilidade de um Classic Load Balancer

1. Antes de alterar as zonas de disponibilidade do balanceador de carga, é uma boa ideia primeiro atualizar as zonas de disponibilidade do grupo do Auto Scaling para verificar se há disponibilidade para seus tipos de instância nas zonas especificadas.

   Para atualizar as zonas de disponibilidade do seu grupo de Auto Scaling, use o comando a seguir update-auto-scaling-group. Substitua a sub-rede IDs de amostra pela IDs das sub-redes nas zonas de disponibilidade a serem ativadas. As sub-redes especificadas substituem as sub-redes habilitadas anteriormente. Substitua *my-asg* pelo nome do seu grupo.

   ```
   aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
     --vpc-zone-identifier "subnet-41767929,subnet-cb663da2"
   ```

2. Use o describe-auto-scaling-groupscomando a seguir para verificar se as instâncias nas novas sub-redes foram iniciadas. Se as instâncias tiverem sido iniciadas, você verá uma lista das instâncias e seus status. Substitua *my-asg* pelo nome do seu grupo.

   ```
   aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
   ```

3. Use o comando attach-load-balancer-to-subnets a seguir para habilitar uma nova zona de disponibilidade para seu Classic Load Balancer. Substitua o ID de sub-rede de amostra pelo ID da sub-rede para habilitar a zona de disponibilidade. Substitua *my-lb* pelo nome do seu balanceador de carga.

   ```
   aws elb attach-load-balancer-to-subnets --load-balancer-name my-lb \
     --subnets subnet-cb663da2
   ```

   Para desativar uma zona de disponibilidade, use o seguinte comando detach-load-balancer-from-subnets. Substitua o ID de sub-rede de amostra pelo ID da sub-rede para a zona de disponibilidade a ser desabilitada. Substitua *my-lb* pelo nome do seu balanceador de carga.

   ```
   aws elb detach-load-balancer-from-subnets --load-balancer-name my-lb \
     --subnets subnet-8360a9e7
   ```

Desvincular seu grupo-alvo ou Classic Load Balancer

O detach-traffic-sourcescomando a seguir separa um grupo-alvo do seu grupo de Auto Scaling quando você não precisar mais dele.

Para a opção,`--auto-scaling-group-name` substitua *my-asg* pelo nome do seu grupo. Para a opção,`--traffic-sources` substitua o ARN de amostra pelo ARN de um grupo de destino para um Application Load Balancer, Network Load Balancer ou balanceador de carga de gateway.

```
aws autoscaling detach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/1234567890123456"
```

Para desvincular um Classic Load Balancer do seu grupo, especifique as opções `--traffic-sources` e,`--type` como no exemplo a seguir. Substitua *my-classic-load-balancer* pelo nome de um Classic Load Balancer. Para a opção,`--type` especifique um valor de **`elb`**.

```
--traffic-sources "Identifier=my-classic-load-balancer" --type elb
```

Remover as verificações de integridade do Elastic Load Balancing

Para remover as verificações de saúde do Elastic Load Balancing do seu grupo de Auto Scaling, use o comando a update-auto-scaling-groupseguir e **`EC2`** especifique como o valor da opção. `--health-check-type` Substitua *my-asg* pelo nome do seu grupo.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "EC2"
```

Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Comandos legados

Os exemplos a seguir mostram como você pode usar comandos legados da CLI para anexar, desvincular e descrever balanceadores de carga e grupos de destino. Eles permanecem neste documento como referência para todos os clientes que desejam usá-los. Continuamos oferecendo suporte aos comandos antigos da CLI, mas recomendamos que você use os novos comandos “fontes de tráfego” da CLI, que podem anexar e desvincular vários tipos de fontes de tráfego. Você

pode usar os comandos antigos da CLI e os comandos “fontes de tráfego” da CLI no mesmo grupo do Auto Scaling.

Anexar seu grupo de destino ou Classic Load Balancer (legado)

Para anexar um grupo de destino

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling com um grupo-alvo anexado. Especifique o nome do recurso da Amazon (ARN) de um grupo de destino para um Application Load Balancer, Network Load Balancer ou balanceador de carga de gateway.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --target-group-arns "arn:aws:elasticloadbalancing:region:account-id:targetgroup/my-
targets/1234567890123456" \
  --min-size 1 --max-size 5
```

O comando attach-load-balancer-target-groups a seguir anexa um grupo-alvo a um grupo existente do Auto Scaling.

```
aws autoscaling attach-load-balancer-target-groups --auto-scaling-group-name my-asg \
  --target-group-arns "arn:aws:elasticloadbalancing:region:account-id:targetgroup/my-
targets/1234567890123456"
```

Para anexar seu Classic Load Balancer

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling com um Classic Load Balancer anexado.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-configuration-name my-launch-config \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --load-balancer-names "my-load-balancer" \
  --min-size 1 --max-size 5
```

O attach-load-balancerscomando a seguir anexa o Classic Load Balancer especificado a um grupo de Auto Scaling existente.

```
aws autoscaling attach-load-balancers --auto-scaling-group-name my-asg \
  --load-balancer-names my-lb
```

Descrever seu grupo de destino ou Classic Load Balancer (legado)

Para descrever grupos de destino

Para descrever os grupos-alvo associados a um grupo do Auto Scaling, use o comando describe-load-balancer-target-groups. O exemplo a seguir lista os grupos-alvo do*my-asg*.

```
aws autoscaling describe-load-balancer-target-groups --auto-scaling-group-name my-asg
```

Descrever Classic Load Balancers

Para descrever os Classic Load Balancers associados a um grupo de Auto Scaling, use o describe-load-balancerscomando. O exemplo a seguir lista os balanceadores de carga clássicos para*my-asg*.

```
aws autoscaling describe-load-balancers --auto-scaling-group-name my-asg
```

Desvincular seu grupo de destino ou Classic Load Balancer (legado)

Para desanexar um grupo de destino

O comando detach-load-balancer-target-groups a seguir separa um grupo-alvo do seu grupo de Auto Scaling quando você não precisar mais dele.

```
aws autoscaling detach-load-balancer-target-groups --auto-scaling-group-name my-asg \
  --target-group-arns "arn:aws:elasticloadbalancing:region:account-id:targetgroup/my-targets/1234567890123456"
```

Desvincular um Classic Load Balancer

O detach-load-balancerscomando a seguir separa um Classic Load Balancer do seu grupo de Auto Scaling quando você não precisar mais dele.

```
aws autoscaling detach-load-balancers --auto-scaling-group-name my-asg \
  --load-balancer-names my-lb
```

Gerenciar o fluxo de tráfego com um grupo de destino do VPC Lattice

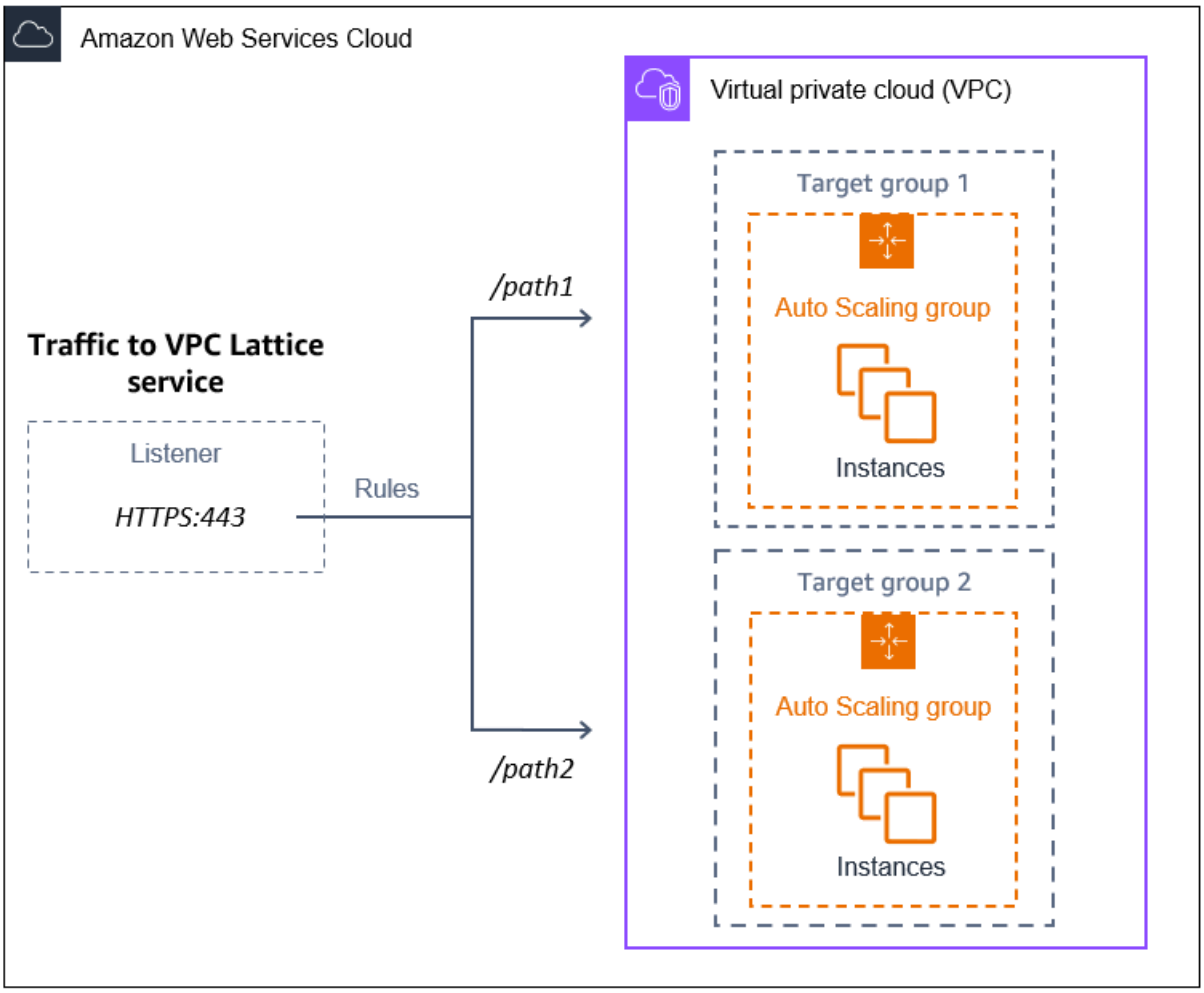
Você pode usar o Amazon VPC Lattice para gerenciar o fluxo de tráfego e as chamadas de API entre suas aplicações e serviços que são executados em recursos separados, como grupos do

Auto Scaling ou funções do Lambda. O VPC Lattice é um serviço de rede de aplicativos que permite conectar, proteger e monitorar todos os seus serviços em várias contas e nuvens privadas virtuais (). VPCs Para saber mais sobre o VPC Lattice, consulte o que é o VPC Lattice?

Para começar a usar o VPC Lattice, primeiro crie os recursos necessários do VPC Lattice que permitem que os recursos em uma VPC associada a uma rede de serviços se conectem entre si. Esses recursos incluem os serviços, os receptores, as regras de receptores e os grupos de destino.

Para associar um grupo do Auto Scaling a um serviço do VPC Lattice, crie um grupo de destino para o serviço que roteie as solicitações para instâncias registradas por ID de instância e adicione um receptor ao serviço que envie solicitações para o grupo de destino. Em seguida, anexe o grupo de destino ao grupo do Auto Scaling. O Amazon EC2 Auto Scaling registra automaticamente as instâncias do EC2 como destinos com o grupo de destino. Posteriormente, quando o Amazon EC2 Auto Scaling precisa encerrar uma instância, ele cancela automaticamente o registro da instância do grupo de destino antes do encerramento.

Depois de anexar o grupo-alvo, ele é o ponto de entrada para todas as solicitações recebidas no seu grupo do Auto Scaling. Como mostra o exemplo no diagrama a seguir, as solicitações recebidas podem então ser roteadas para o grupo-alvo apropriado usando regras do receptor especificadas para um serviço VPC Lattice.



Quando o tráfego é roteado pelo VPC Lattice para seu grupo do Auto Scaling, o VPC Lattice equilibra as solicitações entre as instâncias no grupo usando o balanceamento de carga ida e volta. O VPC Lattice também pode monitorar a integridade das suas instâncias registradas e rotear somente o tráfego para instâncias íntegras.

Para manter suas instâncias disponíveis para solicitações recebidas, você pode, opcionalmente, adicionar verificações de integridade do VPC Lattice ao seu grupo do Auto Scaling. Dessa forma, se uma das instâncias do EC2 falhar, o grupo do Auto Scaling iniciará automaticamente uma nova instância para substituí-la. O comportamento das verificações de integridade do VPC Lattice é semelhante ao comportamento das verificações de integridade do Elastic Load Balancing. As verificações de integridade padrão para um grupo do Auto Scaling são somente verificações de integridade do EC2.

Para saber mais sobre o VPC Lattice, consulte Simplifique a Service-to-Service conectividade, a segurança e o monitoramento com o Amazon VPC Lattice — agora disponível ao público em geral no blog. AWS

Conteúdo

- Preparar para anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling
- Anexe um grupo de destino VPC Lattice ao seu grupo do Auto Scaling
- Verifique o status do anexo do grupo de destino do VPC Lattice

Preparar para anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling

Antes de anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling, será necessário concluir os pré-requisitos a seguir:

- Você já deve ter criado uma rede de serviços, um serviço, um receptor e um grupo-alvo do VPC Lattice. Para obter mais informações, consulte os tópicos a seguir no Guia do usuário do VPC Lattice.
  - Redes de serviços
  - Serviços
  - Listeners
  - Grupos de destino
- O grupo-alvo deve estar na mesma Conta da AWS VPC e região do seu grupo de Auto Scaling.
- O grupo de destino deve especificar um tipo de destino `instance`. Não é possível especificar um tipo de destino `ip` ao usar um grupo do Auto Scaling.
- É necessário ter permissões do IAM suficientes para anexar o grupo de destino ao grupo do Auto Scaling. O exemplo de política a seguir mostra as permissões mínimas necessárias para anexar e separar grupos de destino.

  JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
```

```
            "Action": [
                "autoscaling:AttachTrafficSources",
                "autoscaling:DetachTrafficSources",
                "autoscaling:DescribeTrafficSources",
                "vpc-lattice:RegisterTargets",
                "vpc-lattice:DeregisterTargets"
            ],
            "Resource": "*"
        }
    ]
}
```

- Se o modelo de execução do seu grupo do Auto Scaling não contiver as configurações corretas para o VPC Lattice, como um grupo de segurança compatível, será necessário atualizar o modelo de execução. As instâncias existentes não são atualizadas com as novas configurações quando o modelo de execução é modificado. Para atualizar as instâncias existentes, você pode iniciar uma atualização para substituir a instância. Para obter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling.
- Antes de ativar as verificações de integridade do VPC Lattice no grupo do Auto Scaling, você pode configurar uma verificação de integridade baseada em aplicações para verificar se a aplicação está respondendo conforme o esperado. Para obter mais informações, consulte Verificações de integridade para seus grupos de destino no Guia do usuário do VPC Lattice.

Grupos de segurança: regras de entrada e saída

Os grupos de segurança atuam como firewall para instâncias do EC2 associadas, controlando o tráfego de entrada e saída no nível da instância.

> **Note**
>
> A configuração da rede é suficientemente complexa e recomendamos fortemente que você crie um novo grupo de segurança para uso com o VPC Lattice. Também facilita a ajuda Suporte se você precisar entrar em contato com eles. As seções a seguir se baseiam na suposição de que você segue essa recomendação.
>
> Para saber mais sobre a criação de grupos de segurança para o VPC Lattice que você pode usar com seu grupo do Auto Scaling, consulte Controle o tráfego usando grupos de segurança no Guia do usuário do VPC Lattice. Para solucionar problemas com o fluxo de tráfego, consulte o Guia do usuário do VPC Lattice para obter mais informações.

Para obter informações sobre como criar um grupo de segurança, consulte Criar um grupo de segurança no Manual do Usuário do Amazon EC2 e use a tabela a seguir para determinar quais opções selecionar.

Opção	Valor
Nome	Um nome fácil de lembrar.
Descrição	Uma descrição para ajudar você a identificar o grupo de segurança.
VPC	A mesma VPC do grupo do Auto Scaling.

Regras de entrada

Quando você cria um security group, ele não possui regras de entrada. Nenhum tráfego de entrada originário de clientes em uma rede de serviços do VPC Lattice para a instância será permitido até que você adicione regras de entrada ao grupo de segurança.

Para permitir que clientes em uma rede de serviços do VPC Lattice se conectem às instâncias no grupo do Auto Scaling, o grupo de segurança do grupo do Auto Scaling deve estar configurado corretamente. Nesse caso, defina uma regra de entrada para permitir o tráfego do nome da lista de prefixos AWS gerenciados do VPC Lattice, em vez de um endereço IP específico. A lista de prefixos do VPC Lattice é um intervalo de endereços IP usado pelo VPC Lattice na notação CIDR. Para obter mais informações, consulte Trabalhar com listas AWS de prefixos gerenciadas no Guia do usuário da Amazon VPC.

Para obter informações sobre como adicionar regras a um grupo de segurança, consulte Configuração de regras do grupo de segurança no Guia do usuário da Amazon VPC e use a tabela a seguir para determinar quais opções é necessário selecionar.

Opção	Valor
Regra HTTP	Tipo: HTTP

Opção	Valor	
	Fonte: com.amazonaws.*region*.vpc-lattice	
Regra HTTPS	Tipo: HTTPS Fonte: com.amazonaws.*region*.vpc-lattice	

O grupo de segurança é com estado: permite o tráfego de clientes dentro da rede de serviços do VPC Lattice para instâncias no grupo do Auto Scaling e envia a resposta de volta para o cliente que saiu anteriormente.

Regras de saída

Por padrão, um security group inclui uma regra de saída que permite todo o tráfego de saída. Opcionalmente, você pode remover essa regra padrão e adicionar uma regra de saída para acomodar necessidades específicas de segurança.

Limitações

- Não há suporte a grupos de instâncias mistas. Se você vincular um grupo de destino do VPC Lattice a um grupo do Auto Scaling que tenha uma política de instâncias mistas, o algoritmo de balanceamento de carga distribuirá uniformemente a carga em todos os recursos disponíveis, pressupondo que as instâncias sejam semelhantes o suficiente para lidar com cargas iguais.

Anexe um grupo de destino VPC Lattice ao seu grupo do Auto Scaling

Este tópico descreve como anexar um grupo de destino do VPC Lattice a um grupo do Auto Scaling. Também descreve como ativar as verificações de integridade do VPC Lattice para permitir que o Amazon EC2 Auto Scaling substitua instâncias que o VPC Lattice relata como não íntegras.

Por padrão, o Amazon EC2 Auto Scaling substitui somente instâncias que não sejam íntegras ou inacessíveis com base nas verificações de integridade do Amazon EC2. Se você ativar as verificações de integridade do VPC Lattice, o Amazon EC2 Auto Scaling poderá substituir uma instância em execução se algum dos grupos de destino do VPC Lattice que você anexar ao grupo do Auto Scaling relatar que ela não é íntegra. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

> **Important**
>
> Antes de continuar, preencha todos os pré-requisitos na seção anterior.

Anexar um grupo de destino do VPC Lattice

Você pode anexar um ou vários grupos de destino a um grupo do Auto Scaling ao criar ou atualizar o grupo.

Console

Siga as etapas nesta seção para usar o console para:

- Anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling
- Ative as verificações de integridade do VPC Lattice

Para anexar um grupo de destino do VPC Lattice a um novo grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha Região da AWS onde você criou seu grupo-alvo.
3. Selecione Criar grupo do Auto Scaling.
4. Nas etapas 1 e 2, escolha as opções conforme desejado e prossiga para a Etapa 3: Configurar opções avançadas.
5. Para opções de integração com o VPC Lattice, escolha Anexar ao serviço do VPC Lattice.
6. Em Escolher o grupo de destino do VPC Lattice, escolha seu grupo de destino.
7. (Opcional) Em Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do VPC Lattice.
8. (Opcional) Em Período de carência para verificação de integridade, insira o tempo, em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

9. Prossiga para criar o grupo do Auto Scaling. Suas instâncias serão registradas automaticamente no grupo de destino VPC Lattice após a criação do grupo do Auto Scaling.

Para anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling existente

Use o procedimento a seguir para anexar um grupo de destino de um serviço a um grupo do Auto Scaling existente.

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Detalhes, escolha as opções de integração do VPC Lattice, Editar.
4. Sob as opções de integração com o VPC Lattice, escolha Anexar ao serviço do VPC Lattice.
5. Em Escolher o grupo de destino do VPC Lattice, escolha seu grupo de destino.
6. Selecione Atualizar.

Ao terminar de anexar o grupo de destino, você pode, opcionalmente, ativar as verificações de integridade que o usam.

Para ativar as verificações de integridade da VPC Lattice

1. Na guia Detalhes, escolha Verificações de integridade, Editar.
2. Em Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do VPC Lattice.
3. Em Período de carência da verificação de integridade, insira o tempo, em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.
4. Selecione Atualizar.

AWS CLI

Siga as etapas desta seção para usar o AWS CLI para:

- Anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling
- Ative as verificações de integridade do VPC Lattice

Anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling

Use o create-auto-scaling-groupcomando a seguir para criar um grupo de Auto Scaling e anexar simultaneamente um grupo-alvo do VPC Lattice especificando seu Amazon Resource Name (ARN).

Substitua os valores de amostra por,`--auto-scaling-group-name`,`--vpc-zone-identifier` `--min-size` e `--max-size`. Na opção,`--launch-template` substitua *`my-launch-template`* e *`1`* pelo nome e versão do modelo de execução que você criou para instâncias registradas em um grupo de destino do VPC Lattice. Na opção,`--traffic-sources` substitua o ARN de amostra pelo ARN do seu grupo de destino do VPC Lattice.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --min-size 1 --max-size 5 \
  --traffic-sources "Identifier=arn:aws:vpc-lattice:region:account-id:targetgroup/
tg-0e2f2665eEXAMPLE"
```

Use o attach-traffic-sourcescomando a seguir para anexar um grupo-alvo do VPC Lattice a um grupo do Auto Scaling depois que ele já tiver sido criado.

```
aws autoscaling attach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:vpc-lattice:region:account-id:targetgroup/
tg-0e2f2665eEXAMPLE"
```

Para ativar as verificações de integridade do VPC Lattice

Se você configurou uma verificação de integridade baseada em aplicações para seu grupo de destino do VPC Lattice, é possível ativar essas verificações de integridade. Use o update-auto-scaling-groupcomando create-auto-scaling-groupou com a `--health-check-type` opção e um valor de**`VPC_LATTICE`**. Para especificar o período de carência para as verificações de integridade realizadas pelo seu grupo do Auto Scaling, inclua a opção `--health-check-grace-period` e forneça seu valor em segundos.

```
--health-check-type "VPC_LATTICE" --health-check-grace-period 60
```

Desanexar um grupo de destino do VPC Lattice

Se o VPC Lattice não for mais necessário, use o procedimento a seguir para desanexar o grupo de destino do grupo do Auto Scaling.

Console

Siga as etapas nesta seção para usar o console para:

- Desanexar um grupo de destino do VPC Lattice de um grupo do Auto Scaling
- Desative as verificações de integridade do VPC Lattice

Para desanexar um grupo de destino do VPC Lattice de um grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/e escolha Auto Scaling Groups no painel de navegação.
2. Marque a caixa de seleção ao lado de um grupo existente.

   Um painel dividido é aberto na parte inferior da página.
3. Na guia Detalhes, escolha as opções de integração do VPC Lattice, Editar.
4. Em Opções de integração do VPC Lattice, escolha o ícone de exclusão (X) próximo ao grupo de destino.
5. Selecione Atualizar.

Ao terminar de desanexar o grupo de destino, você pode desativar as verificações de integridade do VPC Lattice.

Para desativar as verificações de integridade do VPC Lattice

1. Na guia Detalhes, escolha Verificações de integridade, Editar.
2. Em Verificações de integridade, Tipos adicionais de verificação de integridade, desmarque Ativar verificações de integridade do VPC Lattice.
3. Escolha Atualizar.

AWS CLI

Siga as etapas desta seção para usar o AWS CLI para:

- Desanexar um grupo de destino do VPC Lattice de um grupo do Auto Scaling
- Desative as verificações de integridade do VPC Lattice

Use o detach-traffic-sourcescomando para separar um grupo-alvo do seu grupo de Auto Scaling quando você não precisar mais dele.

```
aws autoscaling detach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:vpc-lattice:region:account-id:targetgroup/
tg-0e2f2665eEXAMPLE"
```

Para atualizar as verificações de saúde em um grupo do Auto Scaling para que ele não use mais as verificações de saúde do VPC Lattice, use o comando. update-auto-scaling-group Inclua a opção `--health-check-type` e um valor de **`EC2`**.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "EC2"
```

Verifique o status do anexo do grupo de destino do VPC Lattice

Após anexar um grupo de destino do VPC Lattice a um grupo do Auto Scaling, ele entra no estado `Adding` ao registrar as instâncias no grupo. Após todas as instâncias do grupo são registradas, ele entra no estado `Added`. Depois que pelo menos uma instância registrada passa nas verificações de integridade, ele entra no estado `InService`. Após o grupo de destino entrar no estado,`InService` o Amazon EC2 Auto Scaling pode encerrar e substituir todas as instâncias relatadas como não íntegras. Se nenhuma instância registrada passar nas verificações de integridade (por exemplo, devido a um erro na configuração da verificação de integridade), o grupo de destino não entrará no estado `InService`. O Amazon EC2 Auto Scaling não termina e substitui as instâncias.

Quando você desanexa um grupo de destino para um serviço, ele entra no estado `Removing` ao cancelar o registro das instâncias do grupo. As instâncias permanecem em execução após o cancelamento do registro. Por padrão, a drenagem da conexão (atraso de cancelamento de registro) é habilitada. Se a drenagem da conexão estiver habilitada, o VPC Lattice aguardará a conclusão das solicitações em andamento ou a expiração do tempo limite máximo (o que ocorrer primeiro) antes de cancelar o registro das instâncias.

Você pode verificar o status do anexo usando o AWS Command Line Interface (AWS CLI) ou AWS SDKs. Você não pode verificar o status do anexo no console.

Para usar o AWS CLI para verificar o status do anexo

O describe-traffic-sourcescomando a seguir retorna o status do anexo de todas as fontes de tráfego do grupo de Auto Scaling especificado.

```
aws autoscaling describe-traffic-sources --auto-scaling-group-name my-asg
```

O exemplo retorna o ARN do grupo-alvo do VPC Lattice que está anexado ao grupo do Auto Scaling, junto com o status do anexo do grupo de destino no elemento `State`.

```
{
    "TrafficSources": [
        {
            "Identifier": "arn:aws:vpc-lattice:region:account-
id:targetgroup/tg-0e2f2665eEXAMPLE",
            "State": "InService",
            "Type": "vpc-lattice"
        }
    ]
}
```

Use EventBridge para lidar com eventos do Auto Scaling

A Amazon EventBridge, anteriormente chamada de CloudWatch Eventos, ajuda você a configurar regras orientadas por eventos que monitoram recursos e iniciam ações direcionadas que usam outros serviços. AWS

Os eventos do Amazon EC2 Auto Scaling são EventBridge entregues quase em tempo real. Você pode estabelecer EventBridge regras que invocam ações programáticas e notificações em resposta a uma variedade desses eventos. Por exemplo, enquanto as instâncias estão em processo de inicialização ou encerramento, você pode invocar uma AWS Lambda função para realizar uma tarefa pré-configurada.

Os alvos das EventBridge regras podem incluir AWS Lambda funções, tópicos do Amazon SNS, destinos de API, barramentos de eventos e muito mais. Contas da AWS Para obter informações sobre as metas suportadas, consulte as EventBridge metas da Amazon no Guia EventBridge do usuário da Amazon.

Comece criando EventBridge regras com um exemplo usando um tópico e uma EventBridge regra do Amazon SNS. Em seguida, quando um usuário iniciar uma atualização de instância, o Amazon

SNS notificará você por e-mail sempre que um ponto de verificação for alcançado. Para obter mais informações, consulte Crie EventBridge regras, por exemplo, eventos de atualização.

Conteúdo

- Referência de eventos do Amazon EC2 Auto Scaling
- Exemplos de eventos e padrões de atualização de instâncias
- Exemplos de eventos e padrões de pool ativo
- Use EventBridge as regras da Amazon para automatizar ações

Referência de eventos do Amazon EC2 Auto Scaling

Usando a Amazon EventBridge, você pode criar regras que correspondam aos eventos recebidos e encaminhá-los aos alvos para processamento.

Conteúdo

- Eventos de ação do ciclo de vida
- Eventos de escalabilidade bem-sucedidos
- Eventos de escalabilidade sem êxito

Eventos de ação do ciclo de vida

Quando você adiciona ganchos de ciclo de vida ao seu grupo de Auto Scaling, o Amazon EC2 Auto Scaling envia eventos EventBridge para quando uma instância passa para um estado de espera. Os eventos são emitidos com base no melhor esforço.

Tipos de eventos

- Expandir ação do ciclo de vida
- Reduzir ação do ciclo de vida

Expandir ação do ciclo de vida

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling moveu uma instância para um `Pending:Wait` estado devido a um gancho do ciclo de vida de inicialização.

```
{
  "version": "0",
```

```
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-launch Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "87654321-4321-4321-4321-210987654321",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
    "NotificationMetadata": "additional-info",
    "Origin": "EC2",
    "Destination": "AutoScalingGroup"
  }
}
```

Reduzir ação do ciclo de vida

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling moveu uma instância para um `Terminating:Wait` estado devido a um hook de encerramento do ciclo de vida.

> **Important**
>
> Quando um grupo do Auto Scaling retorna instâncias para um pool ativo em redução, o retorno de instâncias para o grupo de aquecimento também pode gerar eventos `EC2 Instance-terminate Lifecycle Action`. Eventos que são entregues quando uma instância passa para o estado de espera em redução têm `WarmPool` como valor para `Destination`. Para obter mais informações, consulte Instance reuse policy.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-terminate Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
```

```
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken":"87654321-4321-4321-4321-210987654321",
    "AutoScalingGroupName":"my-asg",
    "LifecycleHookName":"my-lifecycle-hook",
    "EC2InstanceId":"i-1234567890abcdef0",
    "LifecycleTransition":"autoscaling:EC2_INSTANCE_TERMINATING",
    "NotificationMetadata":"additional-info",
    "Origin": "AutoScalingGroup",
    "Destination": "EC2"
  }
}
```

Eventos de escalabilidade bem-sucedidos

Os exemplos a seguir mostram os tipos de eventos para eventos de escalabilidade bem-sucedidos. Os eventos são emitidos com base no melhor esforço.

Tipos de eventos

- Evento de expansão bem-sucedido
- Evento de redução bem-sucedido

Evento de expansão bem-sucedido

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling executou uma instância com êxito.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance Launch Successful",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
    "instance-arn"
```

```
  ],
  "detail": {
    "StatusCode": "InProgress",
    "Description": "Launching a new EC2 instance: i-12345678",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "EC2",
    "Destination": "AutoScalingGroup"
  }
}
```

Evento de redução bem-sucedido

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling encerrou uma instância com êxito.

> **Important**
>
> Quando um grupo do Auto Scaling retorna instâncias para um pool ativo em redução, o retorno de instâncias para o pool ativo também pode gerar eventos `EC2 Instance Terminate Successful`. Os eventos que são entregues quando uma instância retorna com sucesso ao pool ativo têm `WarmPool` como valor para `Destination`. Para obter mais informações, consulte Instance reuse policy.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance Terminate Successful",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
```

```
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
    "instance-arn"
  ],
  "detail": {
    "StatusCode": "InProgress",
    "Description": "Terminating EC2 instance: i-12345678",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "AutoScalingGroup",
    "Destination": "EC2"
  }
}
```

Eventos de escalabilidade sem êxito

Os exemplos a seguir mostram os tipos de eventos para eventos de escalabilidade malsucedidos. Os eventos são emitidos com base no melhor esforço.

Tipos de eventos

- Evento de expansão sem êxito
- Evento de redução sem êxito

Evento de expansão sem êxito

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling falhou ao iniciar uma instância.

```
{
  "version": "0",
```

```
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance Launch Unsuccessful",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
    "instance-arn"
  ],
  "detail": {
    "StatusCode": "Failed",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "message-text",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "EC2",
    "Destination": "AutoScalingGroup"
  }
}
```

Evento de redução sem êxito

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling falhou ao encerrar uma instância.

> **Important**
>
> Quando um grupo do Auto Scaling retorna instâncias para um pool ativo em redução, deixar de retornar as instâncias ao pool ativo também pode gerar eventos`EC2 Instance Terminate Unsuccessful` . Os eventos que são entregues quando uma instância falha ao retornar ao pool ativo têm `WarmPool` como valor`Destination`. Para obter mais informações, consulte Instance reuse policy.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance Terminate Unsuccessful",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
    "instance-arn"
  ],
  "detail": {
    "StatusCode": "Failed",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "message-text",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "AutoScalingGroup",
    "Destination": "EC2"
  }
}
```

Exemplos de eventos e padrões de atualização de instâncias

Usando a Amazon EventBridge, você pode criar regras que correspondam aos eventos recebidos e encaminhá-los aos alvos para processamento.

Conteúdo

- Eventos de atualização de instância
- A atualização da instância substitui os eventos do ciclo de vida do volume raiz

Eventos de atualização de instância

Os seguintes exemplos mostram eventos do recurso de atualização de instância. Os eventos são emitidos com base no melhor esforço.

Tipos de eventos

- Ponto de verificação alcançado
- Atualização de instância iniciada
- Atualização de instância bem-sucedida
- Falha na atualização de instância
- Atualização de instância cancelada
- Início da reversão da atualização da instância
- A reversão da atualização da instância foi bem-sucedida
- Falha na reversão da atualização da instância

Ponto de verificação alcançado

Quando o número de instâncias substituídas atinge o limite percentual definido para o ponto de verificação, o Amazon EC2 Auto Scaling envia o evento a seguir.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Checkpoint Reached",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "ab00cf8f-9126-4f3c-8010-dbb8cad6fb86",
    "AutoScalingGroupName": "my-asg",
    "CheckpointPercentage": "50",
    "CheckpointDelay": "300"
  }
}
```

Atualização de instância iniciada

Quando o status de uma atualização de instância muda para, o Amazon EC2 Auto Scaling envia o seguinte eventoInProgress.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Started",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

Atualização de instância bem-sucedida

Quando o status de uma atualização de instância muda para, o Amazon EC2 Auto Scaling envia o seguinte eventoSuccessful.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Succeeded",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
```

```
}
```

Falha na atualização de instância

Quando o status de uma atualização de instância muda para, o Amazon EC2 Auto Scaling envia o seguinte eventoFailed.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Failed",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

Atualização de instância cancelada

Quando o status de uma atualização de instância muda para, o Amazon EC2 Auto Scaling envia o seguinte eventoCancelled.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Cancelled",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
```

```
    "AutoScalingGroupName": "my-asg"
  }
}
```

Início da reversão da atualização da instância

Quando o status de uma atualização de instância muda para, o Amazon EC2 Auto Scaling envia o seguinte eventoRollbackInProgress.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Rollback Started",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

A reversão da atualização da instância foi bem-sucedida

Quando o status de uma atualização de instância muda para, o Amazon EC2 Auto Scaling envia o seguinte eventoRollbackSuccessful.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Rollback Succeeded",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
```

```
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

Falha na reversão da atualização da instância

Quando o status de uma atualização de instância muda para, o Amazon EC2 Auto Scaling envia o seguinte evento`Failed`.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Rollback Failed",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

A atualização da instância substitui os eventos do ciclo de vida do volume raiz

Quando você adiciona ganchos de ciclo de vida ao seu grupo de Auto Scaling, o Amazon EC2 Auto Scaling envia eventos EventBridge para quando uma instância passa para um estado de espera. Para obter mais informações, consulte Use ganchos de ciclo de vida para substituir o volume raiz.

Os exemplos a seguir mostram eventos em que seu grupo de Auto Scaling tem instâncias passando por uma substituição do volume raiz durante uma atualização da instância. Os eventos são emitidos com base no melhor esforço.

Exemplos de evento

- Comece a substituir a ação do ciclo de vida do volume raiz
- Finalizar a substituição da ação do ciclo de vida do volume raiz

- Exemplo de padrões de eventos

Comece a substituir a ação do ciclo de vida do volume raiz

Quando uma instância é selecionada para substituição do volume raiz durante a atualização da instância e seu estado muda para`ReplacingRootVolume:Wait`.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-terminate Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "2021-01-13T00:12:37.214Z",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "71514b9d-6a40-4b26-8523-05e7eEXAMPLE",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-terminate-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_TERMINATING",
    "NotificationMetadata": "additional-info",
    "Origin": "AutoScalingGroup",
    "Destination": "AutoScalingGroup",
    "Action": "ReplaceRootVolume"
  }
}
```

Finalizar a substituição da ação do ciclo de vida do volume raiz

Quando uma instância conclui a substituição do volume raiz e seu estado muda para`Pending:Wait`.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-launch Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
```

```
  "time": "2021-01-13T00:12:37.214Z",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "71514b9d-6a40-4b26-8523-05e7eEXAMPLE",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-launch-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
    "NotificationMetadata": "additional-info",
    "Origin": "AutoScalingGroup",
    "Destination": "AutoScalingGroup",
    "Action": "ReplaceRootVolume"
  }
}
```

Exemplo de padrões de eventos

A seção anterior fornece exemplos de eventos emitidos pelo Amazon EC2 Auto Scaling durante uma atualização de instância com a estratégia de substituição do volume raiz. EventBridge os padrões de eventos têm a mesma estrutura dos eventos aos quais eles correspondem. O padrão menciona os campos com os quais você deseja fazer a correspondência e fornece os valores que você está procurando. Os seguintes campos no evento formam o padrão de evento definido na regra para invocar uma ação:

`"source": "aws.autoscaling"`

Identifica que o evento é do Amazon EC2 Auto Scaling.

`"detail-type": "EC2 Instance-launch Lifecycle Action"`

Identifica o tipo de evento.

`"Action": "ReplaceRootVolume"`

Identifica qual ação está sendo executada na instância.

Use o exemplo de padrão de evento a seguir para capturar todos os eventos associados às instâncias que se preparam para substituir seus volumes raiz:

```
{
```

```
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-terminate Lifecycle Action" ],
  "detail": {
      "Action": [ "ReplaceRootVolume" ]
   }
}
```

Use o exemplo de padrão de evento a seguir para capturar todos os eventos associados às instâncias que retornam ao serviço após concluir a substituição do volume raiz:

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
  "detail": {
      "Action": [ "ReplaceRootVolume" ]
   }
}
```

Exemplos de eventos e padrões de pool ativo

O Amazon EC2 Auto Scaling suporta vários padrões predefinidos na Amazon. EventBridge Isso simplifica a forma como um padrão de evento é criado. Você seleciona valores de campo em um formulário e EventBridge gera o padrão para você. No momento, o Amazon EC2 Auto Scaling não oferece suporte a padrões predefinidos para eventos emitidos por um grupo do Auto Scaling com um pool ativo. Você deve inserir o padrão como um objeto JSON. Esta seção e o tópico Crie EventBridge regras para eventos em piscinas aquecidas mostram como usar um padrão de evento para selecionar eventos e enviá-los para destinos.

Para criar EventBridge regras que filtrem os eventos relacionados ao pool aquecido para os quais o Amazon EC2 Auto Scaling envia, EventBridge inclua os campos `Destination` e `Origin` da seção do `detail` evento.

Os valores de `Origin` e `Destination` podem ser:

`EC2` | `AutoScalingGroup` | `WarmPool`

Conteúdo

- Eventos de exemplo
- Exemplo de padrões de eventos

Eventos de exemplo

Quando você adiciona ganchos de ciclo de vida ao seu grupo de Auto Scaling, o Amazon EC2 Auto Scaling envia eventos EventBridge para quando uma instância passa para um estado de espera. Para obter mais informações, consulte Usar ganchos do ciclo de vida com um grupo dinâmico no grupo do Auto Scaling.

Esta seção inclui exemplos desses eventos quando seu grupo do Auto Scaling tem um pool ativo. Os eventos são emitidos com base no melhor esforço.

> **Note**
>
> Para eventos para os quais o Amazon EC2 Auto Scaling EventBridge envia quando a escalabilidade é bem-sucedida, consulte. Eventos de escalabilidade bem-sucedidos Para eventos em que a escalabilidade não é bem-sucedida, consulteEventos de escalabilidade sem êxito .

Exemplos de evento

- Expandir ação do ciclo de vida
- Reduzir ação do ciclo de vida

Expandir ação do ciclo de vida

Os eventos que são entregues quando uma instância faz a transição para um estado de espera por eventos de expansão têm `EC2 Instance-launch Lifecycle Action` como valor`detail-type` . No objeto `detail`, os valores dos atributos `Origin` e `Destination` mostram de onde a instância vem e para onde está indo.

Neste exemplo de evento de expansão, uma nova instância é iniciada e seu estado muda para `Warmed:Pending:Wait` porque ela foi adicionada ao pool ativo. Para obter mais informações, consulteTransições de estado do ciclo de vida para instâncias em um pool ativo.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-launch Lifecycle Action",
  "source": "aws.autoscaling",
```

```
  "account": "123456789012",
  "time": "2021-01-13T00:12:37.214Z",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "71514b9d-6a40-4b26-8523-05e7eEXAMPLE",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-launch-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
    "NotificationMetadata": "additional-info",
    "Origin": "EC2",
    "Destination": "WarmPool"
  }
}
```

Neste exemplo de evento de expansão, o estado da instância muda para `Pending:Wait` porque ela foi adicionada ao grupo do Auto Scaling a partir do pool ativo. Para obter mais informações, consulte Transições de estado do ciclo de vida para instâncias em um pool ativo.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-launch Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "2021-01-19T00:35:52.359Z",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "19cc4d4a-e450-4d1c-b448-0de67EXAMPLE",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-launch-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
    "NotificationMetadata": "additional-info",
    "Origin": "WarmPool",
    "Destination": "AutoScalingGroup"
  }
```

```
}
```

Reduzir ação do ciclo de vida

Os eventos que são entregues quando uma instância faz a transição para um estado de espera em eventos de redução têm `EC2 Instance-terminate Lifecycle Action` como valor para `detail-type`. No objeto `detail`, os valores dos atributos `Origin` e `Destination` mostram de onde a instância vem e para onde está indo.

Neste evento de exemplo, o estado de uma instância muda para `Warmed:Pending:Wait` quando ela é retornada ao pool ativo. Para obter mais informações, consulte Transições de estado do ciclo de vida para instâncias em um pool ativo.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-terminate Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "2022-03-28T00:12:37.214Z",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "42694b3d-4b70-6a62-8523-09a1eEXAMPLE",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-termination-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_TERMINATING",
    "NotificationMetadata": "additional-info",
    "Origin": "AutoScalingGroup",
    "Destination": "WarmPool"
  }
}
```

Exemplo de padrões de eventos

A seção anterior fornece exemplos de eventos emitidos pelo Amazon EC2 Auto Scaling.

EventBridge os padrões de eventos têm a mesma estrutura dos eventos aos quais eles correspondem. O padrão menciona os campos com os quais você deseja fazer a correspondência e fornece os valores que você está procurando.

Os seguintes campos no evento formam o padrão de evento definido na regra para invocar uma ação:

`"source": "aws.autoscaling"`

Identifica que o evento é do Amazon EC2 Auto Scaling.

`"detail-type": "EC2 Instance-launch Lifecycle Action"`

Identifica o tipo de evento.

`"Origin": "EC2"`

Identifica a origem da instância.

`"Destination": "WarmPool"`

Identifica o destino da instância.

Use o padrão de evento de exemplo a seguir para capturar todos os `EC2 Instance-launch Lifecycle Action` eventos associados a instâncias que entram no pool ativo.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
  "detail": {
      "Origin": [ "EC2" ],
      "Destination": [ "WarmPool" ]
   }
}
```

Use o padrão de evento de exemplo a seguir para capturar todos os `EC2 Instance-launch Lifecycle Action` eventos associados a instâncias que saem do pool ativo devido a um evento de expansão.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
```

```
  "detail": {
      "Origin": [ "WarmPool" ],
      "Destination": [ "AutoScalingGroup" ]
   }
}
```

Use o padrão de evento de exemplo a seguir para capturar todos os `EC2 Instance-launch Lifecycle Action` eventos associados a instâncias que são iniciadas diretamente no grupo do Auto Scaling.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
  "detail": {
      "Origin": [ "EC2" ],
      "Destination": [ "AutoScalingGroup" ]
   }
}
```

Use o padrão de evento de exemplo a seguir para capturar todos os `EC2 Instance-terminate Lifecycle Action` eventos associados a instâncias que retornam ao pool ativo ao reduzir a escala vertical.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-terminate Lifecycle Action" ],
  "detail": {
      "Origin": [ "AutoScalingGroup" ],
      "Destination": [ "WarmPool" ]
   }
}
```

Use o exemplo de padrão de evento a seguir para capturar todos os eventos associados a `EC2 Instance-launch Lifecycle Action`, independentemente da origem ou do destino.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ]
}
```

Use EventBridge as regras da Amazon para automatizar ações

Quando um evento é emitido pelo Amazon EC2 Auto Scaling, uma notificação de evento é enviada para a EventBridge Amazon como um arquivo JSON. Você pode escrever uma EventBridge regra para automatizar as ações a serem tomadas quando um padrão de evento corresponder à regra. Se EventBridge detectar um padrão de evento que corresponda a um padrão definido em uma regra, EventBridge invoca o alvo (ou alvos) especificado na regra.

Você pode usar os procedimentos de exemplo nesta seção como ponto de partida.

Talvez a documentação a seguir também seja útil.

- Para executar ações personalizadas em instâncias conforme elas estão sendo iniciadas ou antes que sejam encerradas usando uma função do Lambda, consulte Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda.
- Para invocar uma função Lambda em chamadas de API registradas CloudTrail, consulte Tutorial: AWS Registrar chamadas EventBridge de API usando o Amazon EventBridge User Guide.
- Para obter mais informações sobre como criar regras de eventos, consulte Criação de EventBridge regras da Amazon que reagem a eventos no Guia EventBridge do usuário da Amazon.

Tópicos

- Crie EventBridge regras, por exemplo, eventos de atualização
- Crie EventBridge regras para eventos em piscinas aquecidas

Crie EventBridge regras, por exemplo, eventos de atualização

O exemplo a seguir cria uma EventBridge regra para enviar uma notificação por e-mail. Ele faz isso sempre que seu grupo do Auto Scaling emite um evento quando um ponto de verificação é atingido durante uma atualização de instância. O procedimento para configurar notificações por e-mail usando o Amazon SNS está incluído. Para usar o Amazon SNS para enviar notificações por e-mail, você deve primeiro criar um tópico e, em seguida, assinar seus endereços de e-mail para o tópico.

Para obter mais informações sobre o recurso de atualização de instância, consulte Use uma atualização de instância para atualizar instâncias em um grupo do Auto Scaling.

Crie um tópico do Amazon SNS

Um tópico do SNS é um ponto de acesso lógico, um canal de comunicação que seu grupo do Auto Scaling usa para enviar notificações. Você cria um tópico especificando um nome para o tópico.

Os nomes de tópico devem atender aos seguintes requisitos:

- Ter de 1 a 256 caracteres
- Conter letras maiúsculas e minúsculas ASCIIs, números, sublinhados ou hífens

Para obter mais informações, consulte Criação de um tópico do Amazon SNS no Guia do desenvolvedor do Amazon Simple Notification Service.

Assinar o tópico do Amazon SNS

Para receber as notificações que seu grupo do Auto Scaling envia ao tópico, você deve assinar um endpoint para o tópico. Neste procedimento, em Endpoint, especifique o endereço de e-mail no qual você deseja receber as notificações do Amazon EC2 Auto Scaling.

Para obter instruções, consulte Assinatura de um tópico do Amazon SNS no Guia do desenvolvedor do Amazon Simple Notification Service.

Confirmar sua assinatura do Amazon SNS

O Amazon SNS envia um e-mail de confirmação para o endereço de e-mail especificado na etapa anterior.

Certifique-se de abrir o e-mail em AWS Notificações e escolher o link para confirmar a assinatura antes de continuar com a próxima etapa.

Você receberá uma mensagem de confirmação de. AWS O Amazon SNS agora está configurado para receber notificações e enviar a notificação como um e-mail para o endereço de e-mail que você especificou.

Encaminhar eventos para seu tópico do Amazon SNS

Crie uma regra que corresponda aos eventos selecionados e os encaminhe para o tópico do Amazon SNS para notificar os endereços de e-mail inscritos.

Para criar uma regra que envie notificações para o tópico do Amazon SNS

1. Abra o EventBridge console da Amazon em https://console.aws.amazon.com/events/.

2. No painel de navegação, escolha Regras.
3. Escolha Create rule.
4. Em Define rule detail (Definir detalhe da regra), faça o seguinte:
   a. Informe um Name (Nome) para a regra e, opcionalmente, uma descrição.

      Uma regra não pode ter o mesmo nome que outra regra na mesma região e no mesmo barramento de eventos.
   b. Em Barramento de eventos, escolha padrão. Quando um AWS serviço na sua conta gera um evento, ele sempre vai para o ônibus de eventos padrão da sua conta.
   c. Em Rule type, escolha Rule with an event pattern.
   d. Escolha Próximo.
5. Em Build event pattern (Criar padrão de evento), faça o seguinte:
   a. Em Origem do evento, escolha AWS eventos ou eventos de EventBridge parceiros.
   b. Em Event pattern (Padrão de evento), faça o seguinte:
      i. Para Origem do evento, escolha Serviços da AWS.
      ii. Em AWS service (Serviço da AWS), escolha Auto Scaling.
      iii. Em Event type (Tipo de evento), escolha Instance Refresh (Atualização de instância).
      iv. Por padrão, a regra corresponde a qualquer instância na região. Para criar uma regra que notifique você quando um ponto de verificação for atingido durante uma atualização de instância, escolha Specific instance event(s) (Eventos específicos de instância) e selecione EC2 Auto Scaling Instance Refresh Checkpoint Reached (Ponto de verificação de atualização de instância do EC2 Auto Scaling atingido).
      v. Por padrão, a regra corresponde a qualquer grupo do Auto Scaling na região. Para fazer com que a regra corresponda a um grupo do Auto Scaling específico, escolha Specific group name(s) (Nomes de grupos específicos) e selecione um ou mais grupos do Auto Scaling.
      vi. Escolha Próximo.
6. Em Select target(s) (Selecionar destino(s)), faça o seguinte:
   a. Em Target types (Tipos de destino), escolha AWS service (Serviço da AWS).
   b. Em Select a target (Selecionar um destino), escolha SNS topic (Tópico do SNS).
   c. Em Topic (Tópico), escolha o tópico do Amazon SNS.

d. (Opcional) Em Additional settings (Configurações adicionais), é possível, opcionalmente, definir configurações adicionais. Para obter mais informações, consulte Criação de EventBridge regras da Amazon que reagem a eventos no Guia EventBridge do usuário da Amazon.

e. Escolha Próximo.

7. (Opcional) Em Tags (Etiquetas), é possível atribuir, opcionalmente, uma ou mais etiquetas à sua regra e, em seguida, escolher Next (Próximo).
8. Em Review and create (Revisar e criar), revise os detalhes da regra e modifique-os conforme necessário. Em seguida, escolha Create rule (Criar regra).

Crie EventBridge regras para eventos em piscinas aquecidas

O exemplo a seguir cria uma EventBridge regra para invocar ações programáticas. Ele faz isso sempre que seu grupo do Auto Scaling emitir um evento quando uma nova instância for adicionada ao pool ativo.

Antes de criar a regra, crie a AWS Lambda função que você deseja que a regra use como destino. Você deve especificar essa função como o destino da regra. O procedimento a seguir fornece somente as etapas para criar a EventBridge regra que atua quando novas instâncias entram no pool aquecido. Para obter um tutorial introdutório que mostra como criar uma função simples do Lambda a ser invocada quando um evento recebido corresponder a uma regra, consulte Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda.

Para obter mais informações sobre como criar e trabalhar com pools ativos, consulte Diminua a latência para aplicações com tempos de inicialização longos usando pools ativos.

Para criar uma regra de evento para invocar uma função do Lambda

1. Abra o EventBridge console da Amazon em https://console.aws.amazon.com/events/.
2. No painel de navegação, escolha Regras.
3. Escolha Create rule.
4. Em Define rule detail (Definir detalhe da regra), faça o seguinte:

   a. Informe um Name (Nome) para a regra e, opcionalmente, uma descrição.

      Uma regra não pode ter o mesmo nome que outra regra na mesma região e no mesmo barramento de eventos.

b. Em Barramento de eventos, escolha padrão. Quando um AWS service (Serviço da AWS) em sua conta gera um evento, ele sempre vai para o ônibus de eventos padrão da sua conta.

c. Em Rule type, escolha Rule with an event pattern.

d. Escolha Próximo.

5. Em Build event pattern (Criar padrão de evento), faça o seguinte:

a. Em Origem do evento, escolha AWS eventos ou eventos de EventBridge parceiros.

b. Em Padrão de evento, escolha Padrão personalizado (editor JSON) e cole o padrão a seguir na caixa Padrão de evento, substituindo o texto pelo nome do seu grupo de Auto Scaling. *italics*

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
  "detail": {
      "AutoScalingGroupName": [ "my-asg" ],
      "Origin": [ "EC2" ],
      "Destination": [ "WarmPool" ]
   }
}
```

Para criar uma regra que faça a correspondência com outros eventos, modifique o padrão de evento. Para obter mais informações, consulte Exemplo de padrões de eventos.

c. Escolha Próximo.

6. Em Select target(s) (Selecionar destino(s)), faça o seguinte:

a. Em Target types (Tipos de destino), escolha AWS service (Serviço da AWS).

b. Em Select a target (Selecionar um destino), escolha Lambda function (Função do Lambda).

c. Em Function (Função), escolha a função para a qual deseja enviar os eventos.

d. (Opcional) Em Configure version/alias (Configurar versão/alias), insira as configurações de versão e alias para a função do Lambda de destino.

e. (Opcional) Em Additional settings (Configurações adicionais), insira qualquer configuração adicional conforme adequado para seu aplicativo. Para obter mais informações, consulte Criação de EventBridge regras da Amazon que reagem a eventos no Guia EventBridge do usuário da Amazon.

f. Escolha Próximo.

7. (Opcional) Em Tags (Etiquetas), é possível atribuir, opcionalmente, uma ou mais etiquetas à sua regra e, em seguida, escolher Next (Próximo).
8. Em Review and create (Revisar e criar), revise os detalhes da regra e modifique-os conforme necessário. Em seguida, escolha Create rule (Criar regra).

Fornecer conectividade de rede para suas instâncias do Auto Scaling usando a Amazon VPC

A Amazon Virtual Private Cloud (Amazon VPC) é um serviço que permite lançar AWS recursos como grupos de Auto Scaling em uma rede virtual logicamente isolada que você define.

Uma sub-rede na Amazon VPC é uma subdivisão em uma zona de disponibilidade definida por um segmento de intervalo de endereços IP da VPC. Usando sub-redes, você pode agrupar suas instâncias com base em suas necessidades operacionais e de segurança. Uma sub-rede reside totalmente dentro da Zona de disponibilidade em que foi criada. Você ativa instâncias do Auto Scaling dentro das sub-redes.

Para habilitar a comunicação entre a internet e as instâncias em suas sub-redes, é necessário criar um gateway de internet e anexá-lo à sua VPC. Um gateway de internet permite que seus recursos nas sub-redes conectem-se à internet por meio da borda da rede do Amazon EC2. Se o tráfego de uma sub-rede for roteado para um gateway da Internet, a sub-rede será conhecida como uma sub-rede pública. Se o tráfego de uma sub-rede não for roteado para um gateway de internet, a sub-rede será conhecida como uma sub-rede privada. Use uma sub-rede pública para recursos que devem ser conectados à internet, e uma sub-rede privada para recursos que não precisam ser conectados à internet. Para obter mais informações sobre como fornecer acesso à Internet a instâncias em uma VPC, consulte Acesso à Internet no Guia do usuário da Amazon VPC.

Conteúdo

- VPC padrão
- VPC não padrão
- Considerações sobre a escolha de sub-redes da VPC
- Endereçamento IP em uma VPC
- Interfaces de rede em uma VPC
- Locação de localização de instância

- AWS Outposts
- Mais recursos para aprender sobre VPCs

VPC padrão

Se você criou o seu Conta da AWS depois de 4 de dezembro de 2013 ou está criando seu grupo de Auto Scaling em um novo Região da AWS, criamos uma VPC padrão para você. Sua VPC padrão é fornecida com uma sub-rede padrão em cada Zona de disponibilidade. Por padrão, se você tiver uma VPC padrão, seu grupo do Auto Scaling será criado na VPC padrão.

Você pode visualizá-lo VPCs na VPCs página Sua do console Amazon VPC.

Para obter mais informações sobre a VPC padrão, consulte Padrão VPCs no Guia do usuário da Amazon VPC.

VPC não padrão

Você pode optar por criar mais VPCs acessando a página Painel da VPC em Console de gerenciamento da AWS e selecionando Criar VPC.

Para saber mais, consulte o Manual do usuário da Amazon VPC.

> **Note**
>
> Uma VPC abrange todas as zonas de disponibilidade na Região da AWS. Quando você adicionar sub-redes à VPC, escolha várias zonas de disponibilidade para garantir que as aplicação hospedadas nessas sub-redes sejam altamente disponíveis. Uma zona de disponibilidade é um ou mais datacenters discretos com energia, redes e conectividade redundantes em uma Região da AWS. As zonas de disponibilidade permitem tornar as aplicação em produção altamente disponíveis, tolerantes a falhas e escaláveis.

Considerações sobre a escolha de sub-redes da VPC

Observe os seguintes fatores ao escolher as sub-redes da VPC para seu grupo do Auto Scaling:

- Se você estiver conectando um balanceador de carga Elastic Load Balancing ao seu grupo do Auto Scaling, as instâncias poderão ser iniciadas em sub-redes públicas ou privadas. No entanto,

o balanceador de carga deve ser criado somente nas sub-redes para suportar a resolução de DNS.

- Se você estiver acessando suas instâncias do Auto Scaling diretamente por meio do SSH, as instâncias poderão ser iniciadas somente em sub-redes públicas.
- Se você estiver acessando instâncias do Auto Scaling sem entrada AWS Systems Manager usando o Session Manager, as instâncias podem ser executadas em sub-redes públicas ou privadas.
- Se você estiver usando sub-redes privadas, poderá permitir que as instâncias do Auto Scaling acessem a Internet usando um gateway NAT público.
- Por padrão, as sub-redes padrão em uma VPC padrão são sub-redes públicas.

Endereçamento IP em uma VPC

Quando você ativa suas instâncias do Auto Scaling em uma VPC, um endereço IP privado no intervalo de CIDR da sub-rede na qual a instância foi executada é automaticamente atribuído às suas instâncias. Isso permite que suas instâncias se comuniquem com outras instâncias na VPC.

Você pode configurar um modelo de execução ou uma configuração de execução para atribuir IPv4 endereços públicos às suas instâncias. A atribuição de endereços IP públicos às suas instâncias permite que elas se comuniquem com a Internet ou com outros AWS serviços.

Quando você executa instâncias em uma sub-rede configurada para atribuir IPv6 endereços automaticamente, elas recebem IPv6 endereços IPv4 e endereços. Caso contrário, eles receberão somente IPv4 endereços. Para obter mais informações, consulte IPv6os endereços no Guia do usuário do Amazon EC2.

Para obter informações sobre como especificar intervalos CIDR para a VPC ou sub-rede, consulte o Manual do usuário da Amazon VPC.

O Amazon EC2 Auto Scaling pode atribuir automaticamente endereços IP privados adicionais na execução da instância quando você utiliza um modelo de execução que especifica interfaces de rede adicionais. Cada interface de rede recebe um único endereço IP privado do intervalo CIDR da sub-rede em que a instância é executada. Nesse caso, o sistema não pode mais atribuir automaticamente um IPv4 endereço público à interface de rede primária. Você não poderá se conectar às suas instâncias por meio de um IPv4 endereço público, a menos que associe os endereços IP elásticos disponíveis às instâncias do Auto Scaling.

Interfaces de rede em uma VPC

Cada instância na sua VPC tem uma interface de rede padrão (a interface de rede primária). Você não pode desvincular uma interface de rede primária de uma instância. Você pode criar e anexar uma interface de rede adicional para qualquer instância da VPC. O número de interfaces de rede que você pode anexar varia de acordo com o tipo de instância.

Ao iniciar uma instância usando um modelo de execução, você pode especificar interfaces de rede adicionais. No entanto, iniciar uma instância do Auto Scaling com várias interfaces de rede cria automaticamente cada interface na mesma sub-rede da instância. Isso ocorre porque o Amazon EC2 Auto Scaling ignora as sub-redes definidas no modelo de execução em favor do que é especificado no grupo do Auto Scaling. Para obter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.

Se você criar ou conectar duas ou mais interfaces de rede da mesma sub-rede a uma instância, poderá encontrar problemas de rede, como roteamento assimétrico, especialmente em instâncias que usem uma variante de Linux não Amazon. Se você precisar desse tipo de configuração, deverá configurar a interface de rede secundária dentro do sistema operacional. Para ver um exemplo, consulte Como posso fazer minha interface de rede secundária funcionar na minha instância do Ubuntu EC2? no Centro de AWS Conhecimento.

Locação de localização de instância

Por padrão, todas as instâncias na VPC são executadas como instâncias de locação compartilhada. O Amazon EC2 Auto Scaling também oferece suporte a instâncias dedicadas e hosts dedicados. Para obter mais informações, consulte Criar um modelo de execução usando configurações avançadas.

AWS Outposts

AWS Outposts estende uma Amazon VPC de uma AWS região para um posto avançado com os componentes da VPC que são acessíveis na região, incluindo gateways de internet, gateways privados virtuais, Amazon VPC Transit Gateways e VPC endpoints. Um Outpost fica hospedado em uma zona de disponibilidade na região e é uma extensão dessa zona de disponibilidade que você pode usar para resiliência.

Para obter mais informações, consulte o Guia do usuário do AWS Outposts.

Para obter um exemplo de como implantar um grupo do Auto Scaling que atenda ao tráfego de um Application Load Balancer em um Outpost, consulte a seguinte postagem de blog Configuring an Application Load Balancer on AWS Outposts (Configurar um Application Load Balancer no).

Mais recursos para aprender sobre VPCs

Use os tópicos a seguir para saber mais sobre sub-redes VPCs e sub-redes.

- Sub-redes privadas em uma VPC
  - Exemplo: VPC com servidores em sub-redes privadas e NAT
  - Gateways NAT
- Sub-redes públicas em uma VPC
  - Exemplo: VPC para um ambiente de teste
  - Exemplo: VPC para servidores web e de banco de dados
- Sub-redes para seu Application Load Balancer
  - Sub-redes para seu balanceador de carga
- Informações gerais da VPC
  - Guia do usuário da Amazon VPC
  - Connect VPCs usando o emparelhamento de VPC
  - Interfaces de rede elástica
  - Usar endpoints da VPC para conectividade privada

Segurança no Amazon EC2 Auto Scaling

A segurança na nuvem AWS é a maior prioridade. Como AWS cliente, você se beneficia de uma arquitetura de data center e rede criada para atender aos requisitos das organizações mais sensíveis à segurança.

A segurança é uma responsabilidade compartilhada entre você AWS e você. O Modelo de Responsabilidade Compartilhada descreve isso como segurança da nuvem e segurança na nuvem:

- Segurança da nuvem — AWS é responsável por proteger a infraestrutura que executa AWS os serviços na AWS nuvem. AWS também fornece serviços que você pode usar com segurança. Auditores terceirizados testam e verificam regularmente a eficácia de nossa segurança como parte dos AWS programas de de . Para saber mais sobre os programas de conformidade que se aplicam ao Amazon EC2 Auto Scaling AWS , consulte serviços em escopo por programa de AWS conformidade serviços em escopo por de conformidade.
- Segurança na nuvem — Sua responsabilidade é determinada pelo AWS serviço que você usa. Você também é responsável por outros fatores, incluindo a confidencialidade de seus dados, os requisitos da sua empresa e as leis e normas aplicáveis.

Esta documentação ajuda a entender como aplicar o modelo de responsabilidade compartilhada ao usar o Amazon EC2 Auto Scaling. Os tópicos a seguir mostram como configurar o Amazon EC2 Auto Scaling para atender aos seus objetivos de segurança e de compatibilidade. Você também aprende a usar outros AWS serviços que ajudam você a monitorar e proteger seus recursos do Amazon EC2 Auto Scaling.

Tópicos

- Segurança da infraestrutura no Amazon EC2 Auto Scaling
- Resiliência no Amazon EC2 Auto Scaling
- Amazon EC2 Auto Scaling e proteção de dados
- Gerenciamento de identidade e acesso para o Amazon EC2 Auto Scaling
- Validação de compatibilidade do Amazon EC2 Auto Scaling
- Amazon EC2 Auto Scaling e endpoints da VPC da interface

Segurança da infraestrutura no Amazon EC2 Auto Scaling

Como um serviço gerenciado, o Amazon EC2 Auto Scaling é protegido AWS pela segurança de rede global. Para obter informações sobre serviços AWS de segurança e como AWS proteger a infraestrutura, consulte AWS Cloud Security. Para projetar seu AWS ambiente usando as melhores práticas de segurança de infraestrutura, consulte Proteção de infraestrutura no Security Pillar AWS Well-Architected Framework.

Você usa chamadas de API AWS publicadas para acessar o Amazon EC2 Auto Scaling pela rede. Os clientes devem oferecer compatibilidade com:

- Transport Layer Security (TLS). Exigimos TLS 1.2 e recomendamos TLS 1.3.
- Conjuntos de criptografia com perfect forward secrecy (PFS) como DHE (Ephemeral Diffie-Hellman) ou ECDHE (Ephemeral Elliptic Curve Diffie-Hellman). A maioria dos sistemas modernos, como Java 7 e versões posteriores, comporta esses modos.

Também é possível usar um endpoint da nuvem privada virtual (VPC) com o Amazon EC2 Auto Scaling. Os endpoints da VPC de interface permitem que os recursos de sua Amazon VPC usem seus endereços IP privados para acessar o Amazon EC2 Auto Scaling sem se expor à Internet pública. Para obter mais informações, consulte Amazon EC2 Auto Scaling e endpoints da VPC da interface.

Recursos relacionados

Para obter informações sobre os atributos de isolamento do tráfego de serviços fornecidos pelo Amazon EC2, consulte Segurança da infraestrutura no Amazon EC2 no Guia do usuário do Amazon EC2.

Resiliência no Amazon EC2 Auto Scaling

A infraestrutura AWS global é construída em torno Regiões da AWS de zonas de disponibilidade. Regiões da AWS fornecem várias zonas de disponibilidade fisicamente separadas e isoladas, conectadas a redes de baixa latência, alta taxa de transferência e alta redundância. Com as zonas de disponibilidade, é possível projetar e operar aplicações e bancos de dados que automaticamente executam o failover entre as zonas sem interrupção. As zonas de disponibilidade são altamente disponíveis, tolerantes a falhas e escaláveis que uma ou várias infraestruturas de data center tradicionais.

Para obter mais informações sobre zonas de disponibilidade Regiões da AWS e zonas de disponibilidade, consulte Infraestrutura AWS global.

Para se beneficiar da redundância geográfica do design da zona de disponibilidade, faça o seguinte:

- Estenda seu grupo de Auto Scaling em várias zonas de disponibilidade.
- Mantenha pelo menos uma instância em cada zona de disponibilidade.
- Anexe um balanceador de carga para distribuir o tráfego de entrada nas mesmas zonas de disponibilidade. Se você usar um Application Load Balancer, certifique-se de que cada instância do EC2 obtenha uma quantidade semelhante de tráfego, mantendo o balanceamento de carga entre zonas ativado. Isso ajuda a limitar o impacto do aumento da carga nas instâncias existentes durante um evento de failover e resulta em maior resiliência do que sem o balanceamento de carga entre zonas.
- Certifique-se de que as verificações de integridade do Elastic Load Balancing estejam configuradas corretamente e também de que estejam habilitadas no grupo do Auto Scaling. Então, se uma instância falhar em sua verificação de integridade, o Elastic Load Balancing para de enviar tráfego para ela e redireciona o tráfego para instâncias íntegras, enquanto o Amazon EC2 Auto Scaling substitui a instância não íntegra.

O Amazon EC2 Auto Scaling ajuda a manter disponibilidade de aplicações:

- Verifica as instâncias quanto a problemas de integridade e acessibilidade. Quando uma instância se torna não íntegra, ela encerra automaticamente a instância e inicia uma nova.
- Se as políticas de dimensionamento dinâmico estiverem em vigor, dimensiona automaticamente a capacidade de acordo com o tráfego de entrada.
- Detecta problemas na confiabilidade das CloudWatch métricas da Amazon que suportam políticas de escalabilidade e interrompe as atividades de escalabilidade quando métricas confiáveis não estão disponíveis, como quando faltam pontos de dados.
- Tenta manter números equivalentes de instâncias em cada zona de disponibilidade habilitada à medida que seu grupo aumenta.
- Usa zonas de disponibilidade para manter alta disponibilidade. Quando uma zona de disponibilidade se torna não íntegra, o Amazon EC2 Auto Scaling fará o seguinte:
  - Inicia novas instâncias em zonas de disponibilidade diferentes que estejam habilitadas para seu grupo do Auto Scaling.

  - Redistribui as instâncias em todas as zonas de disponibilidade ativadas quando a zona de disponibilidade não íntegra retorna a um estado íntegro.
- Continua tentando iniciar instâncias em outras zonas de disponibilidade habilitadas se uma instância falhar ao iniciar em uma determinada zona de disponibilidade.
- Registra e cancela o registro automaticamente de instâncias com os balanceadores de carga associados ao seu grupo do Auto Scaling. Dessa forma, você não precisa registrar e cancelar o registro de instâncias separadamente.
- As interrupções do plano de controle do APIs serviço Amazon EC2 Auto Scaling não afetarão a escalabilidade dos grupos de Auto Scaling existentes.

Recursos relacionados

Para obter informações sobre os atributos que ajudam a atender às suas necessidades de resiliência de dados fornecidas pelo Amazon EBS, consulte Resiliência no Amazon Elastic Block Store no Guia do usuário do Amazon EBS.

Amazon EC2 Auto Scaling e proteção de dados

O modelo de responsabilidade AWS compartilhada O modelo de se aplica à proteção de dados no Amazon EC2 Auto Scaling. Conforme descrito neste modelo, AWS é responsável por proteger a infraestrutura global que executa todos os Nuvem AWS. Você é responsável por manter o controle sobre o conteúdo hospedado nessa infraestrutura. Você também é responsável pelas tarefas de configuração e gerenciamento de segurança dos Serviços da AWS que usa. Para obter mais informações sobre privacidade de dados, consulte Perguntas frequentes sobre privacidade de dados . Para obter informações sobre proteção de dados na Europa, consulte o Centro de Regulamentação Geral de Proteção de Dados (GDPR).

Para fins de proteção de dados, recomendamos que você proteja Conta da AWS as credenciais e configure usuários individuais com Centro de Identidade do AWS IAM ou AWS Identity and Access Management (IAM). Dessa maneira, cada usuário receberá apenas as permissões necessárias para cumprir suas obrigações de trabalho. Recomendamos também que você proteja seus dados das seguintes formas:

- Use uma autenticação multifator (MFA) com cada conta.
- Use SSL/TLS para se comunicar com AWS os recursos. Exigimos TLS 1.2 e recomendamos TLS 1.3.

- Configure a API e o registro de atividades do usuário com AWS CloudTrail. Para obter informações sobre o uso de CloudTrail trilhas para capturar AWS atividades, consulte Como trabalhar com CloudTrail trilhas no Guia AWS CloudTrail do usuário.
- Use soluções de AWS criptografia, juntamente com todos os controles de segurança padrão Serviços da AWS.
- Use serviços gerenciados de segurança avançada, como o Amazon Macie, que ajuda a localizar e proteger dados sensíveis armazenados no Amazon S3.
- Se você precisar de módulos criptográficos validados pelo FIPS 140-3 ao acessar AWS por meio de uma interface de linha de comando ou de uma API, use um endpoint FIPS. Para saber mais sobre os endpoints FIPS disponíveis, consulte Federal Information Processing Standard (FIPS) 140-3.

É altamente recomendável que nunca sejam colocadas informações confidenciais ou sensíveis, como endereços de e-mail de clientes, em tags ou campos de formato livre, como um campo Nome. Isso inclui quando você trabalha com o Amazon EC2 Auto Scaling Serviços da AWS ou outro usando o console, a AWS CLI API ou os SDKs. AWS Quaisquer dados inseridos em tags ou em campos de texto de formato livre usados para nomes podem ser usados para logs de faturamento ou de diagnóstico. Se você fornecer um URL para um servidor externo, é fortemente recomendável que não sejam incluídas informações de credenciais no URL para validar a solicitação nesse servidor.

Ao iniciar uma instância no Amazon EC2, você tem a opção de passar dados do usuário para a instância para fazer configuração adicional quando as botas da instância. Também recomendamos que você nunca forneça informações confidenciais ou sigilosas nos dados do usuário que serão passados para uma instância.

Use AWS KMS keys para criptografar volumes do Amazon EBS

Você pode configurar seu grupo do Auto Scaling para criptografar dados de volume do Amazon EBS armazenados na nuvem com o AWS KMS keys. O Amazon EC2 Auto AWS Scaling oferece suporte a chaves gerenciadas e gerenciadas pelo cliente para criptografar seus dados. Observe que a opção `KmsKeyId` para especificar uma chave gerenciada pelo cliente não está disponível quando você usa uma configuração de execução. Para especificar sua chave gerenciada pelo cliente, use um modelo de execução. Para obter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling. Para obter informações sobre como criar, armazenar e gerenciar suas chaves de AWS KMS criptografia, consulte o Guia do AWS Key Management Service desenvolvedor.

Você também pode configurar uma chave gerenciada pelo cliente em sua EBS-backed AMI antes de definir o modelo de execução ou a configuração de execução, ou usar criptografia por padrão para impor a criptografia dos novos volumes do EBS e das cópias instantâneas que você criar. Para obter mais informações, consulte Usar criptografia com EBS-backed AMIs no Guia do usuário do Amazon EC2 e Criptografia por padrão no Guia do usuário do Amazon EBS.

> **Note**
>
> Para obter informações sobre como configurar a política de chave necessária para iniciar instâncias do Auto Scaling ao usar uma chave gerenciada pelo cliente para criptografia, consulte Política de AWS KMS chaves necessária para uso com volumes criptografados.

Recursos relacionados

Para obter as diretrizes de proteção de dados fornecidas pelo Amazon EBS, consulte Proteção de dados no Amazon Elastic Block Store no Guia do Usuário do Amazon EBS.

Política de AWS KMS chaves necessária para uso com volumes criptografados

O Amazon EC2 Auto Scaling usa funções vinculadas a serviços para delegar permissões a outras pessoas. Serviços da AWS As funções vinculadas ao serviço Amazon EC2 Auto Scaling são predefinidas e incluem permissões que o Amazon EC2 Auto Scaling exige para ligar para outras pessoas em seu nome. Serviços da AWS As permissões predefinidas também incluem acesso ao seu Chaves gerenciadas pela AWS. No entanto, elas não incluem acesso às chaves gerenciadas pelo cliente, permitindo que você mantenha o controle total sobre essas chaves.

Este tópico descreve como configurar a política de chaves de que você precisa para iniciar instâncias do Auto Scaling ao especificar uma chave gerenciada pelo cliente para a criptografia do Amazon EBS.

> **Note**
>
> O Amazon EC2 Auto Scaling não precisa de autorização adicional para usar a Chave gerenciada pela AWS padrão para proteger os volumes criptografados em sua conta.

Conteúdo

- Visão geral do
- Configurar políticas de chave
- Exemplo 1: seções da política de chaves que permitem acesso à chave gerenciada pelo cliente
- Exemplo 2: seções da política de chaves que permitem acesso entre contas à chave gerenciada pelo cliente
- Edite as principais políticas no AWS KMS console

Visão geral do

O seguinte AWS KMS keys pode ser usado para a criptografia do Amazon EBS quando o Amazon EC2 Auto Scaling inicia instâncias:

- Chave gerenciada pela AWS: uma chave de criptografia em sua conta que é criada por, pertencente a e gerenciada pelo Amazon EBS. Essa é a chave de criptografia padrão para uma nova conta. O Chave gerenciada pela AWS é usado para criptografia, a menos que você especifique uma chave gerenciada pelo cliente.
- Chave gerenciada pelo cliente: uma chave de criptografia personalizada que você cria, possui e gerencia. Para obter mais informações, consulte Criação de chaves Guia do desenvolvedor do AWS Key Management Service .

  Observação: a chave deve ser simétrica. O Amazon EBS não oferece suporte a chaves gerenciadas pelo cliente assimétricas.

Você configura chaves gerenciadas pelo cliente ao criar snapshots criptografados ou um modelo de execução que especifica volumes criptografados ou ao habilitar a criptografia por padrão.

Configurar políticas de chave

Suas chaves do KMS devem ter uma política de chaves que permita que o Amazon EC2 Auto Scaling execute instâncias com volumes do Amazon EBS criptografados com uma chave gerenciada pelo cliente.

Use os exemplos nesta página para configurar uma política de chaves para conceder ao Amazon EC2 Auto Scaling acesso à sua chave gerenciada pelo cliente. Você pode modificar a política de chaves da chave gerenciada pelo cliente no momento em que a chave é criada ou posteriormente.

É necessário adicionar, no mínimo, duas declarações de política à política de chaves para que ela funcione com o Amazon EC2 Auto Scaling.

- A primeira declaração permite que a identidade do IAM especificada no elemento `Principal` use a chave gerenciada pelo cliente diretamente. Inclui permissões para realizar as `DescribeKey` operações AWS KMS `Encrypt` `DecryptReEncrypt*`,`GenerateDataKey*`,, e na chave.
- A segunda declaração permite que a identidade do IAM especificada no `Principal` elemento use a `CreateGrant` operação para gerar concessões que delegam um subconjunto de suas próprias permissões para aqueles Serviços da AWS que estão integrados com AWS KMS ou outro principal. Isso permite que eles usem a chave para criar recursos criptografados em seu nome.

Ao adicionar as novas declarações de política à sua política de chave, não altere as declarações existentes na política.

Para cada um dos exemplos a seguir, os argumentos que devem ser substituídos, como uma ID de chave ou o nome de uma função vinculada ao serviço, são mostrados como. *`user placeholder text`* Na maioria dos casos, você pode substituir o nome do perfil vinculado ao serviço pelo nome de um perfil vinculado ao serviço do Amazon EC2 Auto Scaling.

Para saber mais, consulte os seguintes recursos:

- Para criar uma chave com o AWS CLI, consulte create-key.
- Para atualizar uma política de chaves com o AWS CLI, consulte put-key-policy.
- Para encontrar um nome do recurso da Amazon (ARN) e um ID de chave, consulte Como encontrar o ID de chave e o ARN no Guia do desenvolvedor do AWS Key Management Service .
- Para obter informações sobre os perfis vinculados ao serviço do Amazon EC2 Auto Scaling, consulte Perfis vinculados ao serviço do Amazon EC2 Auto Scaling.
- Para obter informações sobre a criptografia do Amazon EBS e o KMS em geral, consulte a seção Criptografia do Amazon EBS no Guia do usuário do Amazon EBS e no Guia do desenvolvedor do AWS Key Management Service.

Exemplo 1: seções da política de chaves que permitem acesso à chave gerenciada pelo cliente

Adicione as duas declarações de política a seguir à política de chaves da chave gerenciada pelo cliente, substituindo o ARN de exemplo pelo ARN da função vinculada ao serviço apropriada que tem

permissão de acesso à chave. Neste exemplo, as seções da política concedem à função vinculada ao serviço chamada AWSServiceRoleForAutoScaling permissões para usar a chave gerenciada pelo cliente.

```
{
   "Sid": "Allow service-linked role use of the customer managed key",
   "Effect": "Allow",
   "Principal": {
       "AWS": [
           "arn:aws:iam::account-id:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling"
       ]
   },
   "Action": [
       "kms:Encrypt",
       "kms:Decrypt",
       "kms:ReEncrypt*",
       "kms:GenerateDataKey*",
       "kms:DescribeKey"
   ],
   "Resource": "*"
}
```

```
{
   "Sid": "Allow attachment of persistent resources",
   "Effect": "Allow",
   "Principal": {
       "AWS": [
           "arn:aws:iam::account-id:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling"
       ]
   },
   "Action": [
       "kms:CreateGrant"
   ],
   "Resource": "*",
   "Condition": {
       "Bool": {
           "kms:GrantIsForAWSResource": true
       }
    }
}
```

Exemplo 2: seções da política de chaves que permitem acesso entre contas à chave gerenciada pelo cliente

Se você criar uma chave gerenciada pelo cliente em uma conta diferente da conta usada pelo grupo do Auto Scaling, será necessário usar uma concessão em combinação com a política de chaves para permitir o acesso entre contas à chave.

É necessário concluir duas etapas na seguinte ordem:

1. Em primeiro lugar, adicione as duas seguintes instruções de política à política de chaves da chave gerenciada pelo cliente. Substitua o ARN de exemplo pelo ARN da outra conta, certificando-se de *111122223333* substituí-lo pelo ID real da conta na qual você deseja criar Conta da AWS o grupo Auto Scaling. Isso permite que você conceda permissão para que um usuário ou um perfil do IAM na conta especificada crie uma concessão para a chave usando o seguinte comando da CLI. No entanto, isso por si só não dá a nenhum usuário acesso à chave.

```
{
   "Sid": "Allow external account 111122223333 use of the customer managed key",
   "Effect": "Allow",
   "Principal": {
       "AWS": [
           "arn:aws:iam::111122223333:root"
       ]
   },
   "Action": [
       "kms:Encrypt",
       "kms:Decrypt",
       "kms:ReEncrypt*",
       "kms:GenerateDataKey*",
       "kms:DescribeKey"
   ],
   "Resource": "*"
}
```

```
{
   "Sid": "Allow attachment of persistent resources in external
 account 111122223333",
   "Effect": "Allow",
   "Principal": {
       "AWS": [
           "arn:aws:iam::111122223333:root"
```

```
        ]
    },
    "Action": [
        "kms:CreateGrant"
    ],
    "Resource": "*"
}
```

2. Em seguida, usando a conta na qual deseja criar o grupo do Auto Scaling, crie uma concessão que delegue as permissões relevantes para a função adequada vinculada ao serviço. O elemento `Grantee Principal` da concessão é o ARN do perfil vinculado ao serviço apropriado. O `key-id` é o ARN da chave.

   Veja a seguir um exemplo de comando da CLI create-grant que dá à função vinculada ao serviço AWSServiceRoleForAutoScalingnomeada na *111122223333* conta permissões para usar a chave gerenciada pelo cliente na conta. *444455556666*

```
aws kms create-grant \
  --region us-west-2 \
  --key-id arn:aws:kms:us-
west-2:444455556666:key/1a2b3c4d-5e6f-1a2b-3c4d-5e6f1a2b3c4d \
  --grantee-principal arn:aws:iam::111122223333:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling \
  --operations "Encrypt" "Decrypt" "ReEncryptFrom" "ReEncryptTo" "GenerateDataKey"
 "GenerateDataKeyWithoutPlaintext" "DescribeKey" "CreateGrant"
```

   Para que esse comando seja bem-sucedido, o usuário que faz a solicitação deve ter permissões para a ação `CreateGrant`.

   O exemplo de política do IAM a seguir permite que uma identidade do IAM (usuário ou função) na conta *111122223333* crie uma concessão para a chave gerenciada pelo cliente na conta*444455556666*.

   JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Sid": "AllowCreationOfGrantForTheKMSKeyinExternalAccount444455556666",
      "Effect": "Allow",
      "Action": "kms:CreateGrant",
```

```
      "Resource": "arn:aws:kms:us-
west-2:444455556666:key/1a2b3c4d-5e6f-1a2b-3c4d-5e6f1a2b3c4d"
    }
  ]
}
```

Para obter mais informações sobre como criar uma concessão para uma chave do KMS em uma Conta da AWS diferente, consulte Concessões no AWS KMS no Guia do desenvolvedor do AWS Key Management Service .

> **Important**
>
> O nome do perfil vinculado ao serviço especificado como a entidade principal do beneficiário deve ser o nome de um perfil existente. Depois de criar a concessão, para garantir que a concessão permita que o Amazon EC2 Auto Scaling use a chave especificada do KMS, não exclua e recrie o perfil vinculado ao serviço.

Edite as principais políticas no AWS KMS console

Os exemplos nas seções anteriores mostram apenas como adicionar declarações a uma política de chaves, que é apenas uma maneira de alterar uma política de chaves. A maneira mais fácil de alterar uma política de chaves é usar a visualização padrão do AWS KMS console para políticas de chaves e tornar uma identidade (usuário ou função) do IAM um dos principais usuários da política de chaves apropriada. Para obter mais informações, consulte Usando a visualização Console de gerenciamento da AWS padrão no Guia do AWS Key Management Service desenvolvedor.

> **Important**
>
> Tenha cuidado. As declarações de política de visualização padrão do console incluem permissões para realizar AWS KMS `Revoke` operações na chave gerenciada pelo cliente. Se você conceder Conta da AWS acesso a uma chave gerenciada pelo cliente em sua conta e acidentalmente revogar a concessão que lhes deu essa permissão, os usuários externos não poderão mais acessar seus dados criptografados ou a chave que foi usada para criptografar seus dados.

Gerenciamento de identidade e acesso para o Amazon EC2 Auto Scaling

AWS Identity and Access Management (IAM) é uma ferramenta AWS service (Serviço da AWS) que ajuda o administrador a controlar com segurança o acesso aos AWS recursos. Os administradores do IAM controlam quem pode ser autenticado (conectado) e autorizado (ter permissões) para usar os recursos do Amazon EC2 Auto Scaling. O IAM é um AWS service (Serviço da AWS) que você pode usar sem custo adicional.

Para usar o Amazon EC2 Auto Scaling, você Conta da AWS precisa de um e de suas credenciais de segurança para entrar na sua conta. Para obter mais informações, consulte AWS Credenciais de segurança temporárias no Manual do usuário do IAM.

Para concluir a documentação do IAM, consulte o Guia do usuário do IAM.

Controle de acesso

É possível ter credenciais válidas para autenticar suas solicitações. No entanto, a menos que você tenha permissões, não poderá criar nem acessar os recursos do Amazon EC2 Auto Scaling. Por exemplo, é necessário ter permissões para criar grupos do Auto Scaling, iniciar instâncias com modelos de execução e assim por diante.

As seções a seguir apresentam detalhes sobre como um administrador do IAM pode usá-lo para ajudar a proteger seus recursos do Amazon EC2 Auto Scaling controlando quem pode executar ações do Amazon EC2 Auto Scaling.

Recomendamos ler os tópicos do Amazon EC2 primeiro. Consulte Gerenciamento de identidade e acesso para o Amazon EC2 no Manual do usuário do Amazon EC2 para instâncias do Linux. Depois de ler os tópicos desta seção, você terá uma boa noção do que as permissões de controle de acesso do Amazon EC2 oferecem e como elas podem se adequar às permissões de recursos do Amazon EC2 Auto Scaling.

Tópicos

- Como o Amazon EC2 Auto Scaling funciona com o IAM
- Permissões de API para o Amazon EC2 Auto Scaling
- AWS políticas gerenciadas para Amazon EC2 Auto Scaling

- Perfis vinculados ao serviço do Amazon EC2 Auto Scaling
- Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling
- Prevenção do problema "confused deputy" entre serviços
- Controle o uso do modelo de execução do Amazon EC2 em grupos do Auto Scaling
- Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2

Como o Amazon EC2 Auto Scaling funciona com o IAM

Antes de usar o IAM para gerenciar o acesso ao Amazon EC2 Auto Scaling, saiba quais recursos do IAM estão disponíveis para uso com o Amazon EC2 Auto Scaling.

Recursos do IAM que você pode usar com o Amazon EC2 Auto Scaling

Recurso do IAM	Suporte a Amazon EC2 Auto Scaling
Políticas baseadas em identidade	Sim
Políticas baseadas em recurso	Não
Ações de políticas	Sim
Recursos de políticas	Sim
Chaves de condição de política (específicas do serviço)	Sim
ACLs	Não
ABAC (tags em políticas)	Parcial
Credenciais temporárias	Sim
Perfis de serviço	Sim
Perfis vinculados a serviço	Sim

Para obter uma visão de alto nível de como o Amazon EC2 Auto Scaling e Serviços da AWS outros funcionam com a maioria dos recursos do IAM, Serviços da AWS consulte esse trabalho com o IAM no Guia do usuário do IAM.

Políticas baseadas em identidade para o Amazon EC2 Auto Scaling

Compatível com políticas baseadas em identidade: sim

As políticas baseadas em identidade são documentos de políticas de permissões JSON que podem ser anexados a uma identidade, como usuário do IAM, grupo de usuários ou perfil. Essas políticas controlam quais ações os usuários e perfis podem realizar, em quais recursos e em que condições. Para saber como criar uma política baseada em identidade, consulte Definir permissões personalizadas do IAM com as políticas gerenciadas pelo cliente no Guia do Usuário do IAM.

Com as políticas baseadas em identidade do IAM, é possível especificar ações e recursos permitidos ou negados, assim como as condições sob as quais as ações são permitidas ou negadas. Para saber mais sobre todos os elementos que podem ser usados em uma política JSON, consulte Referência de elemento de política JSON do IAM no Guia do usuário do IAM.

Políticas baseadas em recursos no Amazon EC2 Auto Scaling

Compatibilidade com políticas baseadas em recursos: não

Políticas baseadas em recursos são documentos de políticas JSON que você anexa a um recurso. São exemplos de políticas baseadas em recursos as políticas de confiança de perfil do IAM e as políticas de bucket do Amazon S3. Em serviços compatíveis com políticas baseadas em recursos, os administradores de serviço podem usá-las para controlar o acesso a um recurso específico. Para o atributo ao qual a política está anexada, a política define quais ações uma entidade principal especificado pode executar nesse atributo e em que condições. É necessário especificar uma entidade principal em uma política baseada em recursos. Os diretores podem incluir contas, usuários, funções, usuários federados ou. Serviços da AWS

Para permitir o acesso entre contas, é possível especificar uma conta inteira ou as entidades do IAM em outra conta como a entidade principal em uma política baseada em recursos. Consulte mais informações em Acesso a recursos entre contas no IAM no Guia do usuário do IAM.

Ações de políticas para o Amazon EC2 Auto Scaling

Compatível com ações de políticas: sim

Os administradores podem usar políticas AWS JSON para especificar quem tem acesso ao quê. Ou seja, qual entidade principal pode executar ações em quais recursos e em que condições.

O elemento `Action` de uma política JSON descreve as ações que podem ser usadas para permitir ou negar acesso em uma política. Incluem ações em uma política para conceder permissões para executar a operação associada.

Para ver uma lista das ações do Amazon EC2 Auto Scaling, consulte Ações definidas pelo Amazon EC2 Auto Scaling na Referência de autorização do serviço.

As ações de política no Amazon EC2 Auto Scaling usam o seguinte prefixo antes da ação:

```
autoscaling
```

Para especificar várias ações em uma única declaração, separe-as com vírgulas.

```
"Action": [
  "autoscaling:action1",
  "autoscaling:action2"
]
```

Você pode especificar várias ações usando caracteres curinga (*). Por exemplo, para especificar todas as ações que começam com a palavra `Describe`, inclua a seguinte ação:

```
"Action": "autoscaling:Describe*"
```

Recursos de política para o Amazon EC2 Auto Scaling

Compatível com recursos de políticas: sim

Os administradores podem usar políticas AWS JSON para especificar quem tem acesso ao quê. Ou seja, qual entidade principal pode executar ações em quais recursos e em que condições.

O elemento de política JSON `Resource` especifica o objeto ou os objetos aos quais a ação se aplica. Como prática recomendada, especifique um recurso usando seu nome do recurso da Amazon (ARN). Para ações que não oferecem compatibilidade com permissões em nível de recurso, use um curinga (*) para indicar que a instrução se aplica a todos os recursos.

```
"Resource": "*"
```

Você pode usar ARNs para identificar os grupos do Auto Scaling e iniciar as configurações às quais a política do IAM se aplica.

Um grupo do Auto Scaling tem o ARN a seguir.

```
"Resource": "arn:aws:autoscaling:region:account-
id:autoScalingGroup:uuid:autoScalingGroupName/asg-name"
```

Uma configuração de execução tem o seguinte ARN.

```
"Resource": "arn:aws:autoscaling:region:account-
id:launchConfiguration:uuid:launchConfigurationName/lc-name"
```

Para especificar um grupo do Auto Scaling com a ação `CreateAutoScalingGroup`, é necessário substituir o UUID por um curinga (*), conforme mostrado no exemplo a seguir.

```
"Resource": "arn:aws:autoscaling:region:account-
id:autoScalingGroup:*:autoScalingGroupName/asg-name"
```

Para especificar uma configuração de execução com a ação `CreateLaunchConfiguration`, é necessário substituir o UUID por um curinga (*), conforme mostrado no exemplo a seguir.

```
"Resource": "arn:aws:autoscaling:region:account-
id:launchConfiguration:*:launchConfigurationName/lc-name"
```

Para obter mais informações sobre os tipos de recursos do Amazon EC2 Auto Scaling ARNs e seus, consulte Recursos definidos pelo Amazon EC2 Auto Scaling na Referência de autorização de serviço. Para saber com quais ações você pode especificar o ARN de cada recurso, consulte Ações definidas pelo Amazon EC2 Auto Scaling.

> **Note**
>
> Para ver um exemplo de uma política do IAM usada ARNs para controlar o acesso aos grupos do Auto Scaling, consulte. Controlar quais grupos do Auto Scaling podem ser excluídos

Nem todas as ações do Amazon EC2 Auto Scaling oferecem suporte a permissões em nível de recurso. Para ações que não são compatíveis com permissões em nível de recurso, você precisa usar um curinga (*) como o recurso.

As ações do Amazon EC2 Auto Scaling a seguir oferecem suporte a permissões em nível de recurso.

- `DescribeAccountLimits`
- `DescribeAdjustmentTypes`
- `DescribeAutoScalingGroups`
- `DescribeAutoScalingInstances`
- `DescribeAutoScalingNotificationTypes`
- `DescribeInstanceRefreshes`
- `DescribeLaunchConfigurations`
- `DescribeLifecycleHooks`
- `DescribeLifecycleHookTypes`
- `DescribeLoadBalancers`
- `DescribeLoadBalancerTargetGroups`
- `DescribeMetricCollectionTypes`
- `DescribeNotificationConfigurations`
- `DescribePolicies`
- `DescribeScalingActivities`
- `DescribeScalingProcessTypes`
- `DescribeScheduledActions`
- `DescribeTags`
- `DescribeTerminationPolicyTypes`
- `DescribeWarmPool`

Chaves de condição de política do Amazon EC2 Auto Scaling

Compatível com chaves de condição de política específicas de serviço: sim

Os administradores podem usar políticas AWS JSON para especificar quem tem acesso ao quê. Ou seja, qual entidade principal pode executar ações em quais recursos e em que condições.

O elemento `Condition` especifica quando as instruções são executadas com base em critérios definidos. É possível criar expressões condicionais que usem agentes de condição, como “igual a” ou “menor que”, para fazer a condição da política corresponder aos valores na solicitação. Para ver todas as chaves de condição AWS globais, consulte as chaves de contexto de condição AWS global no Guia do usuário do IAM.

O Amazon EC2 Auto Scaling oferece suporte às seguintes chaves de condição que podem ser usadas para controlar o acesso a ações compatíveis e impor a configuração de grupos do Auto Scaling:

- `autoscaling:InstanceTypes`
- `autoscaling:LaunchConfigurationName`
- `autoscaling:LaunchTemplateVersionSpecified`
- `autoscaling:LoadBalancerNames`
- `autoscaling:MaxSize`
- `autoscaling:MinSize`
- `autoscaling:ResourceTag/key-name: tag-value`
- `autoscaling:TargetGroupARNs`
- `autoscaling:VPCZoneIdentifiers`
- `autoscaling:ForceDelete`

As seguintes chaves de condição são específicas para a criação de solicitações de configuração de lançamento:

- `autoscaling:ImageId`
- `autoscaling:InstanceType`
- `autoscaling:MetadataHttpEndpoint`
- `autoscaling:MetadataHttpPutResponseHopLimit`
- `autoscaling:MetadataHttpTokens`
- `autoscaling:SpotPrice`

O Amazon EC2 Auto Scaling também oferece suporte às seguintes chaves de condição globais que você pode usar para definir permissões com base nas tags na solicitação ou presentes no grupo do Auto Scaling. Para obter mais informações, consulte Etiquetar grupos e instâncias do Auto Scaling.

- `aws:RequestTag/key-name: tag-value`
- `aws:ResourceTag/key-name: tag-value`
- `aws:TagKeys: [tag-key, ...]`

Para saber com que ações da API Amazon EC2 Auto Scaling você pode usar uma chave de condição, consulte Ações definidas pelo Amazon EC2 Auto Scaling na Referência de autorização do serviço. Para obter mais informações sobre chaves de condição do Amazon EC2 Auto Scaling, consulte Chaves de condição para Amazon EC2 Auto Scaling.

> **Note**
>
> Para exemplos de políticas do IAM que usam chaves de condição para controlar o acesso a ações compatíveis e impor a configuração de grupos do Auto Scaling, consulte os seguintes recursos:
>
> - Exigir um modelo de execução e um número de versão — Este exemplo impõe que um modelo de execução e o número da versão do modelo de execução devem ser especificados ao criar ou atualizar grupos do Auto Scaling.
> - Controlar o tamanho de grupos do Auto Scaling que podem ser criados — Este exemplo impõe restrições aos valores possíveis das propriedades `MinSize` e `MaxSize` ao criar ou atualizar grupos do Auto Scaling com uma tag específica.
> - Controlar quais políticas de escalabilidade podem ser excluídas — Este exemplo afirma que a exclusão de políticas de escalabilidade é permitida somente para grupos do Auto Scaling sem uma tag específica.

ACLs no Amazon EC2 Auto Scaling

Suportes ACLs: Não

As listas de controle de acesso (ACLs) controlam quais diretores (membros da conta, usuários ou funções) têm permissões para acessar um recurso. ACLs são semelhantes às políticas baseadas em recursos, embora não usem o formato de documento de política JSON.

ABAC com o Amazon EC2 Auto Scaling

Compatível com ABAC (tags em políticas): parcial

O controle de acesso por atributo (ABAC) é uma estratégia de autorização que define permissões com base em atributos chamados de tags. Você pode anexar tags a entidades e AWS recursos do IAM e, em seguida, criar políticas ABAC para permitir operações quando a tag do diretor corresponder à tag no recurso.

Para controlar o acesso baseado em tags, forneça informações sobre as tags no elemento de condição de uma política usando as `aws:ResourceTag/key-name`, `aws:RequestTag/key-name` ou chaves de condição `aws:TagKeys`.

Se um serviço for compatível com as três chaves de condição para cada tipo de recurso, o valor será Sim para o serviço. Se um serviço for compatível com as três chaves de condição somente para alguns tipos de recursos, o valor será Parcial

Para saber mais sobre o ABAC, consulte Definir permissões com autorização do ABAC no Guia do usuário do IAM. Para visualizar um tutorial com etapas para configurar o ABAC, consulte Usar controle de acesso por atributo (ABAC) no Guia do usuário do IAM.

É possível usar o ABAC em recursos compatíveis com tags, mas nem tudo é compatível com tags. Configurações de execução e políticas de escalabilidade não oferecem suporte a tags, mas os grupos do Auto Scaling sim.

Para obter mais informações, consulte Etiquetar grupos e instâncias do Auto Scaling.

Uso de credenciais temporárias com o Amazon EC2 Auto Scaling

Compatível com credenciais temporárias: sim

As credenciais temporárias fornecem acesso de curto prazo aos AWS recursos e são criadas automaticamente quando você usa a federação ou troca de funções. AWS recomenda que você gere credenciais temporárias dinamicamente em vez de usar chaves de acesso de longo prazo. Para ter mais informações, consulte Credenciais de segurança temporárias no IAM e Serviços da Serviços da AWS que funcionam com o IAM no Guia do usuário do IAM.

Perfis de serviço para o Amazon EC2 Auto Scaling

Compatível com perfis de serviço: sim

O perfil de serviço é um perfil do IAM que um serviço assume para executar ações em seu nome. Um administrador do IAM pode criar, modificar e excluir um perfil de serviço do IAM. Para saber mais, consulte Criar um perfil para delegar permissões a um AWS service (Serviço da AWS) no Guia do Usuário do IAM.

Ao criar um gancho do ciclo de vida que notifica um tópico do Amazon SNS ou uma fila do Amazon SQS, é necessário especificar um perfil para permitir que o Amazon EC2 Auto Scaling acesse o Amazon SNS ou o Amazon SQS em seu nome. Use o console do IAM para configurar o perfil de

serviço para o gancho do ciclo de vida. O console ajuda você a criar um perfil com um conjunto suficiente de permissões usando uma política gerenciada. Para obter mais informações, consulte Receba notificações usando o Amazon SNS e Receba notificações usando o Amazon SQS.

Ao criar um grupo de Auto Scaling, você pode, opcionalmente, passar uma função de serviço para permitir que instâncias do Amazon EC2 acessem outras em seu nome. Serviços da AWS O perfil de serviço para instâncias do Amazon EC2 (também chamado de perfil de instância do Amazon EC2 para um modelo de execução ou configuração de execução) é um tipo especial de perfil de serviço que é atribuído a cada instância do EC2 em um grupo do Auto Scaling quando a instância é executada. Você pode usar o console do IAM e AWS CLI criar ou editar essa função de serviço. Para obter mais informações, consulte Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2.

> **Warning**
>
> A alteração das permissões de um perfil de serviço pode interromper a funcionalidade do Amazon EC2 Auto Scaling. Edite perfis de serviço somente quando o Amazon EC2 Auto Scaling fornecer orientação para isso.

Perfis vinculados ao serviço do Amazon EC2 Auto Scaling

Compatibilidade com perfis vinculados a serviços: sim

Uma função vinculada ao serviço é um tipo de função de serviço vinculada a um. AWS service (Serviço da AWS) O serviço pode assumir o perfil de executar uma ação em seu nome. As funções vinculadas ao serviço aparecem em você Conta da AWS e são de propriedade do serviço. Um administrador do IAM pode visualizar, mas não editar as permissões para perfis vinculados ao serviço.

Para obter detalhes sobre como criar ou gerenciar perfis vinculados ao serviço do Amazon EC2 Auto Scaling, consulte Perfis vinculados ao serviço do Amazon EC2 Auto Scaling.

Permissões de API para o Amazon EC2 Auto Scaling

É necessário conceder permissões aos usuários para chamar as ações da API do Amazon EC2 Auto Scaling de que eles precisam, como indicado em Ações de políticas para o Amazon EC2 Auto Scaling. Além disso, para algumas ações do Amazon EC2 Auto Scaling, você deve conceder aos usuários permissão para chamar ações específicas de outros. AWS APIs

Permissões necessárias de outros AWS APIs

Além das permissões da API Amazon EC2 Auto Scaling, os usuários devem ter as seguintes permissões AWS APIs de outras pessoas para realizar com sucesso a ação associada.

Criar um grupo do Auto Scaling (`autoscaling:CreateAutoScalingGroup`)

- `iam:CreateServiceLinkedRole` — Para criar o perfil vinculado ao serviço padrão, caso esse perfil ainda não exista.
- `iam:PassRole` — Para passar um perfil do IAM para o serviço ou para instâncias do EC2 na execução. Necessário quando um perfil vinculado ao serviço não padrão, um perfil do IAM para um gancho do ciclo de vida ou um modelo de execução que especifica um perfil de instância (um contêiner para um perfil do IAM) é fornecido.
- `ec2:RunInstances` — Para iniciar instâncias quando um modelo de execução é fornecido.
- `ec2:CreateTags` — Marcar instâncias e volumes na execução quando um modelo de execução com uma especificação de tag é fornecido.

Criar um gancho do ciclo de vida (`autoscaling:PutLifecycleHook`)

- `iam:PassRole` — Para passar um perfil do IAM para o serviço. Necessário quando um perfil do IAM é fornecido.

Anexar um grupo de destino do VPC Lattice (`autoscaling:AttachTrafficSources`)

- `vpc-lattice:RegisterTargets` — Para registrar automaticamente as instâncias no grupo de destino.

Desanexar um grupo de destino do VPC Lattice (`autoscaling:DetachTrafficSources`)

- `vpc-lattice:DeregisterTargets` — Para cancelar automaticamente o registro de instâncias com o grupo de destino.

Criar uma configuração de execução (`autoscaling:CreateLaunchConfiguration`)

- `ec2:DescribeImages`
- `ec2:DescribeInstances`
- `ec2:DescribeInstanceAttribute`
- `ec2:DescribeKeyPairs`
- `ec2:DescribeSecurityGroups`
- `ec2:DescribeSpotInstanceRequests`
- `ec2:DescribeVpcClassicLink`

- `iam:PassRole` — Para passar um perfil do IAM para instâncias do EC2 na execução. Necessário quando uma configuração de execução especifica um perfil de instância (um contêiner para um perfil do IAM).

AWS políticas gerenciadas para Amazon EC2 Auto Scaling

Uma política AWS gerenciada é uma política autônoma criada e administrada por AWS. AWS as políticas gerenciadas são projetadas para fornecer permissões para muitos casos de uso comuns, para que você possa começar a atribuir permissões a usuários, grupos e funções.

Lembre-se de que as políticas AWS gerenciadas podem não conceder permissões de privilégio mínimo para seus casos de uso específicos porque elas estão disponíveis para uso de todos os AWS clientes. Recomendamos que você reduza ainda mais as permissões definindo as políticas gerenciadas pelo cliente que são específicas para seus casos de uso.

Você não pode alterar as permissões definidas nas políticas AWS gerenciadas. Se AWS atualizar as permissões definidas em uma política AWS gerenciada, a atualização afetará todas as identidades principais (usuários, grupos e funções) às quais a política está anexada. AWS é mais provável que atualize uma política AWS gerenciada quando uma nova AWS service (Serviço da AWS) é lançada ou novas operações de API são disponibilizadas para serviços existentes.

Para saber mais, consulte AWS Políticas gerenciadas pela no Guia do usuário do IAM.

Políticas gerenciadas do Amazon EC2 Auto Scaling

Você pode anexar as seguintes políticas gerenciadas às suas identidades AWS Identity and Access Management (IAM) (usuários ou funções). Cada política fornece acesso a todas ou a algumas das ações de API para o Amazon EC2 Auto Scaling.

- AutoScalingConsoleFullAccess— Concede acesso total ao Amazon EC2 Auto Console de gerenciamento da AWS Scaling usando o. Esta política funciona quando você usa configurações de execução, mas não quando usa modelos de execução.
- AutoScalingConsoleReadOnlyAccess— Concede acesso somente de leitura ao Amazon EC2 Auto Scaling usando o. Console de gerenciamento da AWS Esta política funciona quando você usa configurações de execução, mas não quando usa modelos de execução.
- AutoScalingFullAccess— Concede acesso total ao Amazon EC2 Auto Scaling para identidades do IAM que precisam de acesso total ao Amazon EC2 Auto Scaling a partir do acesso ou, mas AWS CLI não. SDKs Console de gerenciamento da AWS

- AutoScalingReadOnlyAccess— Concede acesso somente de leitura ao Amazon EC2 Auto Scaling para identidades do IAM que estão fazendo chamadas somente para o ou. AWS CLI SDKs

Ao usar modelos de execução via console, você precisa conceder permissões adicionais específicas para os modelos de execução, o que é debatido em Controle o uso do modelo de execução do Amazon EC2 em grupos do Auto Scaling. O console do Amazon EC2 Auto Scaling precisa de permissões para ações do `ec2` para que ele possa exibir informações sobre modelos de execução e iniciar instâncias usando modelos de execução.

AutoScalingServiceRolePolicy AWS política gerenciada

Essa política é anexada a um perfil vinculado ao serviço que permite que o Amazon EC2 Auto Scaling execute ações em seu nome. Para obter mais informações, consulte Perfis vinculados ao serviço do Amazon EC2 Auto Scaling.

Para visualizar as permissões para esta política, consulte AutoScalingServiceRolePolicy na Referência de políticas gerenciadas pela AWS .

Atualizações do Amazon EC2 Auto Scaling AWS para políticas gerenciadas

Veja detalhes sobre as atualizações das políticas AWS gerenciadas do Amazon EC2 Auto Scaling desde que esse serviço começou a rastrear essas alterações. Para receber alertas automáticos sobre mudanças nesta página, assine o RSS feed na página de histórico de documentos do Amazon EC2 Auto Scaling.

Alteração	Descrição	Data
O Amazon EC2 Auto Scaling atualiza as permissões para sua função vinculada ao serviço	A `AutoScalingServiceRolePolicy` política agora inclui a `ec2:CreateReplaceRootVolumeTask` permissão que permite substituir volumes raiz em instâncias como parte de uma atualização de instância. Além disso, a `events:DescribeRule` permissão foi removida da política. Para	18 de novembro de 2025

Alteração	Descrição	Data
	obter mais informações, consulte Perfis vinculados ao serviço do Amazon EC2 Auto Scaling.	
O Amazon EC2 Auto Scaling adiciona permissões à respectiva função vinculada ao serviço	A `AutoScalingServiceRolePolicy` política agora inclui permissão para chamar a ação da AWS Resource Groups ListGroupResourcesAPI para obter todos os nomes de recursos (ARNs) dos recursos que são membros de um grupo de recursos especificado. Para obter mais informações, consulte Perfis vinculados ao serviço do Amazon EC2 Auto Scaling.	20 de novembro de 2024

Alteração	Descrição	Data
O Amazon EC2 Auto Scaling adiciona permissões à respectiva função vinculada ao serviço	A `AutoScalingServiceRolePolicy` política agora concede permissões para chamar a ação da GetSecurityGroupsForVpcAPI do Amazon EC2 para obter todos os grupos de segurança de uma VPC para melhorar a validação, e a ação da GetInstanceTypesFromInstanceRequirementsAPI do Amazon EC2 para obter informações sobre quais tipos de instância atendem a um determinado conjunto de requisitos de instância. Para obter mais informações, consulte Perfis vinculados ao serviço do Amazon EC2 Auto Scaling.	29 de fevereiro de 2024

Alteração	Descrição	Data
O Amazon EC2 Auto Scaling adiciona permissões ao respectivo perfil vinculado ao serviço	A política `AutoScalingServiceRolePolicy` agora concede permissões ao serviço para acessar as ações de API necessárias para uma integração com o VPC Lattice. • Ações `GetTargetGroup` e `ListTargetGroup` . Obrigatório para recuperar informações sobre grupos de destino VPC Lattice. • Ações `RegisterTargets` e `DeregisterTargets` . Obrigatório para registrar e cancelar o registro de instâncias de grupos de destino VPC Lattice. • `ListTargets` . Permite que o Amazon EC2 Auto Scaling recupere informações de integridade para instâncias registradas em grupos de destino VPC Lattice. Para obter mais informações, consulte Perfis vinculados ao serviço do Amazon EC2 Auto Scaling.	6 de dezembro de 2022

Alteração	Descrição	Data
O Amazon EC2 Auto Scaling adiciona permissões à respectiva função vinculada ao serviço	Para oferecer suporte ao uso de um AWS Systems Manager parâmetro como alias para uma ID de AMI ao criar um modelo de lançamento, a `AutoScalingService RolePolicy` política agora concede permissão para chamar a ação da AWS Systems Manager GetParameters API. Para obter mais informações, consulte Perfis vinculados ao serviço do Amazon EC2 Auto Scaling.	28 de março de 2022
O Amazon EC2 Auto Scaling adiciona permissões à respectiva função vinculada ao serviço	Para oferecer suporte à escalabilidade preditiva , a política `AutoScali ngServiceRolePolic y` agora inclui permissão para chamar a ação de API CloudWatch GetMetricData. Para obter mais informações, consulte Perfis vinculados ao serviço do Amazon EC2 Auto Scaling.	19 de maio de 2021
O Amazon EC2 Auto Scaling começou a monitorar alterações	O Amazon EC2 Auto Scaling começou a monitorar as alterações em suas políticas AWS gerenciadas.	19 de maio de 2021

Perfis vinculados ao serviço do Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling usa perfis vinculados ao serviço para as permissões necessárias para chamar outros serviços da Serviços da AWS em seu nome. Uma função vinculada ao serviço é um tipo exclusivo de função do IAM vinculada diretamente a uma. AWS service (Serviço da AWS)

Os perfis vinculados a serviços oferecem uma maneira segura de delegar permissões a outros serviços da Serviços da AWS , pois somente o serviço vinculado pode assumir um perfil vinculado ao serviço. Para ter mais informações, consulte Criar um perfil vinculado ao serviço no Guia do usuário do IAM. As funções vinculadas ao serviço também permitem que todas as chamadas de API sejam visíveis por meio de. AWS CloudTrail Isso ajuda com os requisitos de monitoramento e auditoria porque você pode rastrear todas as ações que o Amazon EC2 Auto Scaling executa em seu nome. Para obter mais informações, consulte Registre chamadas da API Amazon EC2 Auto Scaling com AWS CloudTrail.

As seções a seguir descrevem como criar e gerenciar perfis vinculados ao serviço do Amazon EC2 Auto Scaling. Comece configurando permissões para autorizar uma identidade do IAM (por exemplo, um usuário ou um perfil) a criar, editar ou excluir um perfil vinculado ao serviço.

Conteúdo

- Visão geral do
- Permissões concedidas pelo perfil vinculado ao serviço
- Regiões compatíveis com funções vinculadas ao serviço do Amazon EC2 Auto Scaling
- Criar, editar e excluir um perfil vinculado ao serviço
  - Criar uma função vinculada ao serviço (automática)
  - Criar um perfil vinculado ao serviço (manual)
  - Editar o perfil vinculado ao serviço
  - Excluir o perfil vinculado ao serviço

Visão geral do

Há dois tipos de funções vinculadas ao serviço do Amazon EC2 Auto Scaling:

- A função vinculada ao serviço padrão para sua conta, chamada AWSServiceRoleForAutoScaling. Essa função é automaticamente atribuída aos seus grupos do Auto Scaling, a menos que você especifique outra função vinculada ao serviço.

- Uma função vinculada ao serviço com um sufixo personalizado que você especifica ao criar a função, por exemplo, AWSServiceRoleForAutoScaling_*mysuffix*.

As permissões de um perfil vinculado ao serviço com sufixo personalizado são idênticas às do perfil vinculado ao serviço padrão. Em ambos os casos, você não poderá editar as funções nem excluí-las se elas ainda estiverem em uso por um grupo do Auto Scaling. A única diferença é o sufixo do nome do perfil.

Você pode especificar qualquer uma das funções ao editar suas políticas de AWS Key Management Service chaves para permitir que as instâncias lançadas pelo Amazon EC2 Auto Scaling sejam criptografadas com sua chave gerenciada pelo cliente. No entanto, se você planeja conceder acesso granular a uma determinada CMK gerenciada pelo cliente, será necessário usar um perfil vinculado ao serviço com sufixo personalizado. O uso de um perfil vinculado ao serviço com sufixo personalizado fornece:

- Mais controle sobre a chave gerenciada pelo cliente
- A capacidade de rastrear qual grupo do Auto Scaling fez uma chamada de API em seus registros CloudTrail

Se você criar chaves gerenciadas pelo cliente às quais nem todos os usuários devem ter acesso, siga estas etapas para permitir o uso de um perfil vinculado ao serviço com sufixo personalizado:

1. Crie um perfil vinculado ao serviço com um sufixo personalizado. Para obter mais informações, consulte Criar um perfil vinculado ao serviço (manual).
2. Conceda ao perfil vinculado ao serviço acesso a uma chave gerenciada pelo cliente. Para obter mais informações sobre a política de chaves que permite que a chave seja usada por um perfil vinculado ao serviço, consulte Política de AWS KMS chaves necessária para uso com volumes criptografados.
3. Dê aos usuários acesso ao perfil vinculado ao serviço que você criou. Para obter mais informações sobre como criar políticas do IAM, consulte Controle qual função vinculada ao serviço pode ser passada (usando) PassRole. Se os usuários tentarem especificar um perfil vinculado ao serviço sem permissão para passar esse perfil para o serviço, eles receberão um erro.

Permissões concedidas pelo perfil vinculado ao serviço

O Amazon EC2 Auto Scaling usa o perfil vinculado ao serviço chamado AWSServiceRoleForAutoScaling ou seu perfil vinculado ao serviço de sufixo personalizado.

O perfil vinculado ao serviço confia no seguinte serviço para assumir o perfil:

- `autoscaling.amazonaws.com`

A política de permissões do perfil, AutoScalingServiceRolePolicy, permite que o Amazon EC2 Auto Scaling execute as ações a seguir:

- `ec2` - Criar, descrever, modificar, iniciar/interromper e encerrar instâncias do EC2.
- `iam` - Passar perfis do IAM para instâncias do EC2 para que as aplicações em execução nas instâncias possam acessar credenciais temporárias para o perfil.
- `iam`— Crie a função vinculada ao serviço AWSServiceRoleForEC2Spot para permitir que o Amazon EC2 Auto Scaling lance Instâncias Spot em seu nome.
- `elasticloadbalancing` - Registrar e cancelar o registro de instâncias com o Elastic Load Balancing e verificar a integridade dos alvos registrados.
- `cloudwatch`— crie, descreva, modifique e exclua CloudWatch alarmes para políticas de escalabilidade e recupere métricas usadas para escalabilidade preditiva.
- `sns` - Publicar notificações no Amazon SNS quando as instâncias forem iniciadas ou encerradas.
- `events`— Crie, descreva, atualize e exclua EventBridge regras em seu nome.
- `ssm`: leia os parâmetros do Parameter Store ao usar um parâmetro do Systems Manager como um alias para um ID de AMI em um modelo de execução.
- `vpc-lattice` - Registrar e cancelar o registro de instâncias com o Elastic Load Balancing e verificar a integridade dos alvos registrados.
- `resource-groups`— Obtenha todos os nomes de recursos (ARNs) dos recursos que são membros de um grupo de recursos especificado.

Regiões compatíveis com funções vinculadas ao serviço do Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling oferece suporte ao uso de funções vinculadas a serviços em todos os lugares em que o serviço está Regiões da AWS disponível.

Criar, editar e excluir um perfil vinculado ao serviço

Criar uma função vinculada ao serviço (automática)

O Amazon EC2 Auto Scaling cria a função vinculada ao serviço AWSServiceRoleForAutoScaling para você na primeira vez que você cria um grupo do Auto Scaling, a menos que você crie manualmente uma função vinculada ao serviço com sufixo personalizado e especifique-a ao criar o grupo.

É necessário ter permissões do IAM para criar a função vinculada ao serviço. Caso contrário, a criação automática falhará. Para obter mais informações, consulte Permissões de perfil vinculado ao serviço no Manual do usuário do IAM e Criar um perfil vinculado ao serviço neste guia.

Criar um perfil vinculado ao serviço (manual)

Para criar um perfil vinculado ao serviço (console)

1. Abra o console do IAM em https://console.aws.amazon.com/iam/.
2. No painel de navegação, selecione Perfil e então, Criar perfil.
3. Em Select trusted entity (Selecionar entidade confiável), escolha AWS service (serviço).
4. Em Choose the service that will use this role (Escolha o serviço que usará esse perfil), escolha EC2 Auto Scaling (Auto Scaling do EC2) e o caso de uso EC2 Auto Scaling (Auto Scaling do EC2) .
5. Escolha Next: Permissions (Próximo: permissões), Next: Tags (Próximo: tags) e Next: Review (Próximo: revisão). Observação: você não pode anexar tags a perfis vinculados ao serviço durante a criação.
6. Na página Revisar, deixe o nome da função em branco para criar uma função vinculada ao serviço com o nome AWSServiceRoleForAutoScalingou digite um sufixo para criar uma função vinculada ao serviço com o nome _. AWSServiceRoleForAutoScaling ***suffix***
7. (Opcional) Em Role description (Descrição do perfil), edite a descrição para o perfil vinculado ao serviço.
8. Selecione Criar perfil.

Para criar um perfil vinculado ao serviço (AWS CLI)

Use o seguinte comando da create-service-linked-roleCLI para criar uma função vinculada ao serviço para o Amazon EC2 Auto Scaling com o nome _. AWSServiceRoleForAutoScaling ***suffix***

```
aws iam create-service-linked-role --aws-service-name autoscaling.amazonaws.com --
custom-suffix suffix
```

A saída desse comando inclui o ARN da função vinculada ao serviço, o qual você pode usar para conceder acesso à chave gerenciada pelo cliente para a função vinculada ao serviço.

```
{
    "Role": {
        "RoleId": "ABCDEF0123456789ABCDEF",
        "CreateDate": "2018-08-30T21:59:18Z",
        "RoleName": "AWSServiceRoleForAutoScaling_suffix",
        "Arn": "arn:aws:iam::123456789012:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling_suffix",
        "Path": "/aws-service-role/autoscaling.amazonaws.com/",
        "AssumeRolePolicyDocument": {
            "Version": "2012-10-17",
            "Statement": [
                {
                    "Action": [
                        "sts:AssumeRole"
                    ],
                    "Principal": {
                        "Service": [
                            "autoscaling.amazonaws.com"
                        ]
                    },
                    "Effect": "Allow"
                }
            ]
        }
    }
}
```

Para ter mais informações, consulte Criar um perfil vinculado ao serviço no Guia do usuário do IAM.

Editar o perfil vinculado ao serviço

Você não pode editar os perfis vinculados ao serviço criados para o Amazon EC2 Auto Scaling. Depois de criar um perfil vinculado ao serviço, você não pode alterar o nome do perfil ou suas permissões. No entanto, você poderá editar a descrição do perfil. Para obter mais informações, consulte Editar uma descrição de perfil vinculado ao serviço no Guia do usuário do IAM.

Excluir o perfil vinculado ao serviço

Se você não estiver usando um grupo do Auto Scaling, recomendamos excluir o perfil vinculado ao serviço. Excluir o perfil evita que você tenha uma entidade que não é usada ou mantida e monitorada ativamente.

Você poderá excluir um perfil vinculado ao serviço somente depois de excluir os recursos dependentes relacionados. Isso evita que você revogue acidentalmente as permissões do Amazon EC2 Auto Scaling para seus recursos. Se um perfil vinculado ao serviço é usado com vários grupos do Auto Scaling, é necessário excluir todos os grupos do Auto Scaling que usam o perfil vinculado ao serviço antes de excluí-lo. Para obter mais informações, consulte Excluir infraestrutura do Auto Scaling.

É possível usar o IAM para excluir um perfil vinculado ao serviço. Para obter mais informações, consulte Excluir uma função vinculada ao serviço no Guia do usuário do IAM.

Se você excluir a função vinculada ao serviço AWSServiceRoleForAutoScaling, o Amazon EC2 Auto Scaling criará a função novamente quando você criar um grupo do Auto Scaling sem especificar outra função vinculada ao serviço.

Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling

Por padrão, um novo usuário não Conta da AWS tem permissão para fazer nada. Um administrador do IAM deve criar e atribuir políticas do IAM que concedam a uma identidade do IAM (como um usuário ou perfil) permissão para executar ações de API do Amazon EC2 Auto Scaling.

Para saber como criar uma política do IAM usando esses exemplos de documentos de política JSON, consulte Criar políticas na aba JSON no Manual do usuário do IAM.

A seguir, um exemplo de uma política de permissões.

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
```

```
            "autoscaling:CreateAutoScalingGroup",
            "autoscaling:UpdateAutoScalingGroup",
            "autoscaling:DeleteAutoScalingGroup"
        ],
        "Resource": "*",
        "Condition": {
            "StringEquals": { "autoscaling:ResourceTag/purpose": "testing" }
        }
    },
    {
        "Effect": "Allow",
        "Action": "autoscaling:Describe*",
        "Resource": "*"
    }]
}
```

Este exemplo de política concede permissões para criar, atualizar e excluir grupos do Auto Scaling, mas somente se o grupo usar a etiqueta **`purpose=testing`**. Como as ações `Describe` não oferecem suporte a permissões em nível de recurso, é necessário especificá-las em uma declaração separada sem condições. Para iniciar instâncias com um modelo de execução, o usuário também precisa ter a permissão/`ec2:RunInstances`. Para obter mais informações, consulte Controle o uso do modelo de execução do Amazon EC2 em grupos do Auto Scaling.

> **Note**
>
> É possível criar suas próprias políticas personalizadas do IAM para permitir ou negar permissões para identidades do IAM (usuários ou perfis) para executar ações do Amazon EC2 Auto Scaling. Você pode anexar essas políticas personalizadas às identidades do IAM que exigem as permissões especificadas. Os exemplos a seguir mostram permissões para alguns casos de uso comuns.
>
> Algumas ações de API do Amazon EC2 Auto Scaling permitem incluir grupos do Auto Scaling específicos na política que podem ser criados ou modificados pela ação. Você pode restringir os recursos de destino para essas ações especificando um grupo individual de Auto ARNs Scaling. No entanto, como prática recomendada, sugerimos usar políticas baseadas em tags que permitam (ou neguem) ações em grupos do Auto Scaling com uma tag específica.

Exemplos

- Controlar o tamanho de grupos do Auto Scaling que podem ser criados

- Controlar quais chaves de tag e valores de tag podem ser usados
- Controlar quais grupos do Auto Scaling podem ser excluídos
- Controlar quais políticas de escalabilidade podem ser excluídas
- Controlar o acesso às ações de atualização da instância
- Criar um perfil vinculado ao serviço
- Controle qual função vinculada ao serviço pode ser passada (usando) PassRole

Controlar o tamanho de grupos do Auto Scaling que podem ser criados

A política a seguir concede permissões para criar e atualizar todos os grupos do Auto Scaling com a tag,**`environment=development`** desde que o solicitante não especifique um tamanho mínimo menor que **`1`** ou um tamanho máximo maior que **`10`**. Sempre que possível, use tags para ajudar você a controlar o acesso aos grupos do Auto Scaling na sua conta.

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:CreateAutoScalingGroup",
          "autoscaling:UpdateAutoScalingGroup"
      ],
      "Resource": "*",
      "Condition": {
          "StringEquals": { "autoscaling:ResourceTag/environment":
 "development" },
          "NumericGreaterThanEqualsIfExists": { "autoscaling:MinSize": 1 },
          "NumericLessThanEqualsIfExists": { "autoscaling:MaxSize": 10 }
      }
   }]
}
```

Como alternativa, se você não estiver usando tags para controlar o acesso aos grupos de Auto Scaling, você pode usar ARNs para identificar os grupos de Auto Scaling aos quais a política do IAM se aplica.

Um grupo do Auto Scaling tem o ARN a seguir.

```
"Resource": "arn:aws:autoscaling:region:account-
id:autoScalingGroup:*:autoScalingGroupName/my-asg"
```

Também possível especificar vários ARNs incluindo-os em uma lista. Para obter mais informações sobre como especificar os ARNs dos recursos do Amazon EC2 Auto Scaling no elemento, `Resource` consulte Recursos de política para o Amazon EC2 Auto Scaling.

Controlar quais chaves de tag e valores de tag podem ser usados

Também é possível usar condições adicionais em suas políticas do IAM para controlar as chaves de tag e os valores que podem ser aplicados aos grupos do Auto Scaling. Para conceder permissões para criar ou etiquetar um grupo do Auto Scaling somente se o solicitante especificar determinadas etiquetas, use a chave de condição `aws:RequestTag`. Para permitir somente chaves de tags específicas, use a chave de condição `aws:TagKeys` com o modificador `ForAllValues`.

A política a seguir requer que o solicitante especifique uma etiqueta com a chave **`environment`** na solicitação. O valor **`"?*"`** impõe que haja um valor para a chave de tag. Para usar um caractere curinga, é necessário usar o operador de condição `StringLike`.

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:CreateAutoScalingGroup",
          "autoscaling:CreateOrUpdateTags"
      ],
      "Resource": "*",
      "Condition": {
          "StringLike": { "aws:RequestTag/environment": "?*" }
      }
   }]
}
```

A política a seguir especifica que o solicitante só pode marcar grupos do Auto Scaling com as etiquetas **`purpose=webserver`** e **`cost-center=cc123`** e permite somente as etiquetas **`purpose`** e **`cost-center`** (nenhuma outra etiqueta pode ser especificada).

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:CreateAutoScalingGroup",
          "autoscaling:CreateOrUpdateTags"
      ],
      "Resource": "*",
      "Condition": {
          "StringEquals": {
              "aws:RequestTag/purpose": "webserver",
              "aws:RequestTag/cost-center": "cc123"
          },
          "ForAllValues:StringEquals": { "aws:TagKeys": ["purpose", "cost-
center"] }
      }
   }]
}
```

A política a seguir requer que o solicitante especifique pelo menos uma etiqueta na solicitação e permite somente as chaves **`cost-center`** e **`owner`**.

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:CreateAutoScalingGroup",
          "autoscaling:CreateOrUpdateTags"
      ],
      "Resource": "*",
```

```
        "Condition": {
            "ForAnyValue:StringEquals": { "aws:TagKeys": ["cost-center", "owner"] }
        }
    }]
}
```

> **Note**
>
> Para condições, a chave de condição não diferencia maiúsculas de minúsculas, e o valor da condição diferencia maiúsculas de minúsculas. Portanto, para aplicar a diferenciação de maiúsculas de minúsculas de uma tag, use a chave de condição `aws:TagKeys`, onde a chave da tag é especificada como um valor na condição.

Controlar quais grupos do Auto Scaling podem ser excluídos

A política a seguir permite a exclusão de um grupo do Auto Scaling somente se o grupo tiver a tag **`environment=development`**.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [{
        "Effect": "Allow",
        "Action": "autoscaling:DeleteAutoScalingGroup",
        "Resource": "*",
        "Condition": {
            "StringEquals": { "aws:ResourceTag/environment": "development" }
        }
    }]
}
```

A política a seguir usa a chave de `autoscaling:ForceDelete` condição para controlar o acesso à ação `DeleteAutoScalingGroup` da API. Isso pode impedir que determinados usuários usem a operação de exclusão forçada, que encerra todas as instâncias em um grupo de Auto Scaling ao excluir o grupo.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [{
        "Effect": "Deny",
        "Action": "autoscaling:DeleteAutoScalingGroup",
        "Resource": "*",
        "Condition": {
            "Bool": {
                "autoscaling:ForceDelete": "true"
            }
        }
    }]
}
```

Como alternativa, se você não estiver usando chaves de condição para controlar o acesso aos grupos do Auto Scaling, você pode especificar o ARNs número de recursos no `Resource` elemento para controlar o acesso.

A política a seguir dá aos usuários permissões para usar a ação da `DeleteAutoScalingGroup` API, mas somente para grupos de Auto Scaling cujo nome começa com. **`devteam-`**

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "autoscaling:DeleteAutoScalingGroup",
            "Resource": "arn:aws:autoscaling:us-
east-1:111122223333:autoScalingGroup:*:autoScalingGroupName/devteam-*"
        }
    ]
}
```

Você também pode especificar vários ARNs colocando-os em uma lista. A inclusão da UUID garante que o acesso seja concedido ao grupo do Auto Scaling específico. O UUID para um novo grupo é diferente do UUID para um grupo excluído com o mesmo nome.

```
"Resource": [
    "arn:aws:autoscaling:region:account-
id:autoScalingGroup:uuid:autoScalingGroupName/devteam-1",
    "arn:aws:autoscaling:region:account-
id:autoScalingGroup:uuid:autoScalingGroupName/devteam-2",
    "arn:aws:autoscaling:region:account-
id:autoScalingGroup:uuid:autoScalingGroupName/devteam-3"
]
```

Controlar quais políticas de escalabilidade podem ser excluídas

A política a seguir concede permissões para usar a ação `DeletePolicy` para excluir uma política de escalabilidade. No entanto, ela também negará a ação se o grupo do Auto Scaling que está recebendo a ação tiver a tag **`environment=production`**. Sempre que possível, use tags para ajudar você a controlar o acesso aos grupos do Auto Scaling na sua conta.

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": "autoscaling:DeletePolicy",
      "Resource": "*"
   },
   {
      "Effect": "Deny",
      "Action": "autoscaling:DeletePolicy",
      "Resource": "*",
      "Condition": {
          "StringEquals": { "autoscaling:ResourceTag/environment": "production" }
      }
   }]
}
```

Controlar o acesso às ações de atualização da instância

A política a seguir concede permissão para iniciar, reverter e cancelar uma atualização de instância somente se o grupo do Auto Scaling que está recebendo a ação tiver a tag **`environment=testing`**. Como as ações `Describe` não oferecem suporte a permissões em nível de recurso, é necessário especificá-las em uma declaração separada sem condições.

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:StartInstanceRefresh",
          "autoscaling:CancelInstanceRefresh",
          "autoscaling:RollbackInstanceRefresh"
      ],
      "Resource": "*",
      "Condition": {
          "StringEquals": { "autoscaling:ResourceTag/environment": "testing" }
      }
   },
   {
      "Effect": "Allow",
      "Action": "autoscaling:DescribeInstanceRefreshes",
      "Resource": "*"
   }]
}
```

Para especificar uma configuração desejada na chamada,`StartInstanceRefresh` é possível que os usuários precisem de algumas permissões relacionadas, como:

- ec2: RunInstances — Para iniciar instâncias do EC2 usando um modelo de execução, o usuário deve ter a `ec2:RunInstances` permissão em uma política do IAM. Para obter mais informações, consulte Controle o uso do modelo de execução do Amazon EC2 em grupos do Auto Scaling.
- ec2: CreateTags — Para iniciar instâncias do EC2 a partir de um modelo de execução que adiciona tags às instâncias e volumes na criação, o usuário deve ter a `ec2:CreateTags`

permissão em uma política do IAM. Para obter mais informações, consulte Permissões necessárias para marcar instâncias e volumes.

- iam: PassRole — Para iniciar instâncias do EC2 a partir de um modelo de execução que contém um perfil de instância (um contêiner para uma função do IAM), o usuário também deve ter a `iam:PassRole` permissão em uma política do IAM. Para obter mais informações e um exemplo de política do IAM, consulte Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2.
- ssm: GetParameters — Para iniciar instâncias do EC2 a partir de um modelo de execução que usa um AWS Systems Manager parâmetro, o usuário também deve ter a `ssm:GetParameters` permissão em uma política do IAM. Para obter mais informações, consulte Use AWS Systems Manager parâmetros em vez de AMI IDs em modelos de lançamento.

Criar um perfil vinculado ao serviço

O Amazon EC2 Auto Scaling exige permissões para criar uma função vinculada ao serviço na primeira vez que qualquer usuário Conta da AWS em você chama as ações da API do Amazon EC2 Auto Scaling. Se a função vinculada ao serviço ainda não existir, o Amazon EC2 Auto Scaling a criará em sua conta. A função vinculada ao serviço concede permissões ao Amazon EC2 Auto Scaling para que ele possa ligar para outras pessoas em seu nome. Serviços da AWS

Para que a criação automática do perfil seja bem-sucedida, os usuários devem ter permissões para a ação `iam:CreateServiceLinkedRole`.

```
"Action": "iam:CreateServiceLinkedRole"
```

O exemplo a seguir mostra uma política de permissões que permite que um usuário crie um perfil vinculado ao serviço do Amazon EC2 Auto Scaling para o Amazon EC2 Auto Scaling.

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": "iam:CreateServiceLinkedRole",
      "Resource": "arn:aws:iam::*:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling",
```

```
      "Condition": {
          "StringLike": { "iam:AWSServiceName":"autoscaling.amazonaws.com" }
      }
   }]
}
```

Controle qual função vinculada ao serviço pode ser passada (usando) PassRole

Os usuários que criam ou atualizam grupos do Auto Scaling e especificam um perfil vinculado ao serviço de sufixo personalizado na solicitação necessitam da permissão `iam:PassRole`.

Você pode usar a `iam:PassRole` permissão para proteger a segurança de suas chaves gerenciadas pelo AWS KMS cliente se conceder a diferentes funções vinculadas ao serviço acesso a chaves diferentes. Dependendo das necessidades de sua organização, talvez você tenha uma chave para a equipe de desenvolvimento, outra para a equipe de QA e outra para a equipe financeira. Primeiro, crie uma função vinculada ao serviço com acesso à chave necessária, por exemplo, uma função vinculada ao serviço chamada AWSServiceRoleForAutoScaling_devteamkeyaccess. Em seguida, anexe a política a uma identidade do IAM, como um usuário ou um perfil.

A política a seguir concede permissões para passar o perfil **`AWSServiceRoleForAutoScaling_devteamkeyaccess`** para qualquer grupo do Auto Scaling cujo nome comece com **`devteam-`**. Se a identidade do IAM que cria o grupo do Auto Scaling tentar especificar um perfil vinculado ao serviço diferente, ela receberá um erro. Se optarem por não especificar uma função vinculada a serviço, a função AWSServiceRoleForAutoScaling padrão será usada.

Para obter mais informações sobre perfis vinculados ao serviço com sufixo personalizado, consulte Perfis vinculados ao serviço do Amazon EC2 Auto Scaling.

Prevenção do problema "confused deputy" entre serviços

O problema "confused deputy" é um problema de segurança em que uma entidade que não tem permissão para executar uma ação pode coagir uma entidade mais privilegiada a executar a ação.

Em AWS, a falsificação de identidade entre serviços pode resultar em um problema confuso de delegado. A personificação entre serviços pode ocorrer quando um serviço (o serviço de chamada) chama outro serviço (o serviço chamado). O serviço de chamada pode ser manipulado de modo a usar suas permissões para atuar nos recursos de outro cliente de uma forma na qual ele não deveria ter permissão para acessar.

Para evitar isso, AWS fornece ferramentas que ajudam você a proteger seus dados para todos os serviços com diretores de serviços que receberam acesso aos recursos em sua conta. Recomendamos o uso das chaves de contexto de condição global `aws:SourceArn` e `aws:SourceAccount` nas políticas de confiança para funções do serviço do Amazon EC2 Auto Scaling. Essas chaves limitam as permissões que o Amazon EC2 Auto Scaling concede a outro serviço ao recurso.

Os valores dos `SourceAccount` campos `SourceArn` e são definidos quando o Amazon EC2 Auto AWS Security Token Service Scaling AWS STS usa () para assumir uma função em seu nome.

Para usar as chaves de condição globais `aws:SourceArn` ou `aws:SourceAccount`, defina o valor como o nome do recurso da Amazon (ARN) ou a conta do recurso que o Amazon EC2 Auto Scaling armazena. Sempre que possível, use `aws:SourceArn`, que é mais específico. Defina o valor como o ARN ou um padrão de ARN com curinga (*) para as partes desconhecidas do ARN. Se você não conhece o ARN do recurso, use `aws:SourceAccount` em vez disso.

O exemplo a seguir mostra como é possível usar as chaves de contexto de condição global `aws:SourceArn` e `aws:SourceAccount` no Amazon EC2 Auto Scaling para evitar o problema do substituto confuso.

Exemplo: uso das chaves de condição **`aws:SourceArn`** e **`aws:SourceAccount`**

A função de serviço é uma função que um serviço assume para realizar ações em seu nome. Nos casos em que você quiser criar ganchos de ciclo de vida que enviem notificações para qualquer lugar que não seja a Amazon EventBridge, você deve criar uma função de serviço para permitir que o Amazon EC2 Auto Scaling envie notificações para um tópico do Amazon SNS ou fila do Amazon SQS em seu nome. Se quiser que apenas um grupo do Auto Scaling seja associado ao acesso entre serviços, você pode especificar a política de confiança da do perfil de serviço da seguinte forma.

Este exemplo de política de confiança usa declarações de condição para limitar a capacidade de `AssumeRole` na função de serviço somente para as ações que afetam o grupo do Auto Scaling especificado na conta especificada. As condições `aws:SourceArn` e `aws:SourceAccount` são avaliadas de forma independente. Qualquer solicitação para usar o perfil de serviço deve atender às duas condições.

Antes de usar essa política, substitua os valores de Região, ID da conta, UUID e nome de grupo por valores válidos da sua conta.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": {
        "Sid": "ConfusedDeputyPreventionExamplePolicy",
        "Effect": "Allow",
        "Principal": {
            "Service": "autoscaling.amazonaws.com"
        },
        "Action": "sts:AssumeRole",
        "Condition": {
            "ArnLike": {
                "aws:SourceArn": "arn:aws:autoscaling:us-
east-1:111122223333:autoScalingGroup:uuid:autoScalingGroupName/my-asg"
            },
            "StringEquals": {
                "aws:SourceAccount": "111122223333"
            }
        }
    }
}
```

No exemplo anterior:

- O elemento `Principal` especifica a entidade principal de serviço do serviço (`autoscaling.amazonaws.com`).
- O elemento `Action` especifica a ação`sts:AssumeRole`.
- O elemento `Condition` especifica as chaves de condição globais `aws:SourceArn` e `aws:SourceAccount`. O ARN da fonte inclui o ID da conta, portanto, não é necessário usar `aws:SourceAccount` com `aws:SourceArn`.

Mais informações

Para obter mais informações, consulte Chaves de contexto de condição globais da AWS, O problema do representante confuso e Atualização de uma política de confiança de um perfil no Guia do usuário do IAM.

Controle o uso do modelo de execução do Amazon EC2 em grupos do Auto Scaling

O Amazon EC2 Auto Scaling oferece suporte ao uso de modelos de execução do Amazon EC2 com seus grupos do Auto Scaling. Recomendamos permitir que os usuários criem grupos do Auto Scaling com base em modelos de execução, pois isso permite que eles usem os recursos mais recentes do Amazon EC2 Auto Scaling e Amazon EC2. Por exemplo, os usuários devem especificar um modelo de execução para usar uma política de instâncias mistas.

É possível usar a política `AmazonEC2FullAccess` para conceder aos usuários acesso total para trabalhar com recursos do Amazon EC2 Auto Scaling, modelos de execução e outros recursos do EC2 em suas contas. Ou é possível criar suas próprias políticas personalizadas do IAM para conceder aos usuários permissões refinadas para trabalhar com modelos de execução, conforme descrito neste tópico.

Uma política de exemplo que você pode personalizar para seu próprio uso

O exemplo a seguir mostra uma política de permissões básica que você pode personalizar para seu próprio uso. A política concede permissões para criar, atualizar e excluir todos os grupos do Auto Scaling, mas somente se o grupo usa a tag **`purpose=testing`**. Em seguida, concede permissão para todas as ações `Describe`. Como as ações `Describe` não oferecem suporte a permissões em nível de recurso, é necessário especificá-las em uma declaração separada sem condições.

Identidades do IAM (usuários ou perfis) com esta política têm permissão para criar ou atualizar um grupo do Auto Scaling usando um modelo de execução porque eles também têm permissão para usar a ação `ec2:RunInstances`.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
          {
            "Effect": "Allow",
            "Action": [
                "autoscaling:CreateAutoScalingGroup",
                "autoscaling:UpdateAutoScalingGroup",
                "autoscaling:DeleteAutoScalingGroup"
            ],
```

```
            "Resource": "*",
            "Condition": {
                "StringEquals": { "autoscaling:ResourceTag/purpose": "testing" }
            }
        },
        {
            "Effect": "Allow",
            "Action": [
                "autoscaling:Describe*",
                "ec2:RunInstances"
            ],
            "Resource": "*"
        }
    ]
}
```

Os usuários que criam ou atualizam grupos do Auto Scaling podem precisar de algumas permissões relacionadas, como:

- ec2: CreateTags — Para adicionar tags às instâncias e volumes na criação, o usuário deve ter a `ec2:CreateTags` permissão em uma política do IAM. Para obter mais informações, consulte Permissões necessárias para marcar instâncias e volumes.
- iam: PassRole — Para iniciar instâncias do EC2 a partir de um modelo de execução que contém um perfil de instância (um contêiner para uma função do IAM), o usuário também deve ter a `iam:PassRole` permissão em uma política do IAM. Para obter mais informações e um exemplo de política do IAM, consulte Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2.
- ssm: GetParameters — Para iniciar instâncias do EC2 a partir de um modelo de execução que usa um AWS Systems Manager parâmetro, o usuário também deve ter a `ssm:GetParameters` permissão em uma política do IAM. Para obter mais informações, consulte Use AWS Systems Manager parâmetros em vez de AMI IDs em modelos de lançamento.

Essas permissões para ações a serem concluídas ao iniciar instâncias são verificadas quando o usuário interage com um grupo do Auto Scaling. Para obter mais informações, consulte Validação de permissões para `ec2:RunInstances` e `iam:PassRole`.

Os exemplos a seguir mostram declarações de políticas que você pode usar para controlar os acessos que os usuários do IAM têm para usar modelos de execução.

Tópicos

- Exigir modelos de execução que têm uma tag específica
- Exigir um modelo de execução e um número de versão
- Exigir o uso do serviço de metadados de instância versão 2 () IMDSv2
- Restringir o acesso aos recursos do Amazon EC2
- Permissões necessárias para marcar instâncias e volumes
- Permissões adicionais do modelo de execução
- Validação de permissões para ec2:RunInstances e iam:PassRole
- Recursos relacionados

Exigir modelos de execução que têm uma tag específica

Ao conceder `ec2:RunInstances` permissões, você pode especificar que os usuários só possam usar modelos de execução com tags específicas ou específicos IDs para limitar as permissões ao iniciar instâncias com um modelo de execução. Você também pode controlar a AMI e outros recursos aos quais qualquer pessoa que use modelos de execução possa fazer referência e usar ao iniciar instâncias especificando permissões adicionais em nível de recurso para a chamada`RunInstances`.

O exemplo a seguir restringe permissões à ação `ec2:RunInstances` para iniciar modelos que estão localizados na região especificada e que têm a tag **`purpose=testing`**. Ele também dá aos usuários acesso aos recursos especificados em um modelo de execução: tipos de instância AMIs, volumes, pares de chaves, interfaces de rede e grupos de segurança.

Para obter mais informações sobre o uso de políticas baseadas em tags com modelos de execução, consulte Controlar o acesso aos modelos de execução com permissões do IAM no Guia do usuário do Amazon EC2 para instâncias do Linux.

Exigir um modelo de execução e um número de versão

Você também pode usar as permissões do IAM para obrigar que um modelo de execução e o número da versão do modelo de execução sejam especificados ao criar ou atualizar grupos do Auto Scaling.

O exemplo a seguir permite que os usuários criem e atualizem grupos do Auto Scaling somente se um modelo de execução e o número da versão do modelo de execução forem especificados. Se usuários com essa política omitirem o número da versão para especificar a versão `$Latest` ou `$Default` do modelo de execução ou tentarem usar uma configuração de execução, a ação falhará.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "autoscaling:CreateAutoScalingGroup",
                "autoscaling:UpdateAutoScalingGroup"
            ],
            "Resource": "*",
            "Condition": {
                "Bool": { "autoscaling:LaunchTemplateVersionSpecified": "true" }
            }
        },
        {
            "Effect": "Deny",
            "Action": [
                "autoscaling:CreateAutoScalingGroup",
                "autoscaling:UpdateAutoScalingGroup"
            ],
            "Resource": "*",
            "Condition": {
                "Null": { "autoscaling:LaunchConfigurationName": "false" }
            }
        }
    ]
}
```

Exigir o uso do serviço de metadados de instância versão 2 () IMDSv2

Para maior segurança, você pode definir as permissões de seus usuários para exigir o uso de um modelo de lançamento que exija IMDSv2. Para obter mais informações, consulte Configurar o serviço de metadados de instância no Guia do usuário do Amazon EC2 para instâncias do Linux.

O exemplo a seguir especifica que os usuários não podem chamar a `ec2:RunInstances` ação, a menos que a instância também tenha optado por exigir o uso de IMDSv2 (indicado por`"ec2:MetadataHttpTokens":"required"`).

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "RequireImdsV2",
            "Effect": "Deny",
            "Action": "ec2:RunInstances",
            "Resource": "arn:aws:ec2:*:*:instance/*",
            "Condition": {
                "StringNotEquals": { "ec2:MetadataHttpTokens": "required" }
            }
        }
    ]
}
```

> **Tip**
>
> Para forçar a substituição de instâncias do ajuste de escala automático que usam um novo modelo de execução ou uma nova versão de um modelo de execução com as opções de metadados de instância configuradas, você pode iniciar uma atualização de instâncias. Para obter mais informações, consulte Atualizar instâncias do Auto Scaling.

Restringir o acesso aos recursos do Amazon EC2

Os exemplos a seguir mostram como usar permissões em nível de recurso e recursos com tags para restringir o acesso aos recursos do Amazon EC2.

Exemplo 1: restrição das execuções de instâncias do Amazon EC2 a recursos e tipos de instância específicos

O exemplo a seguir controla a configuração das instâncias que um usuário pode iniciar restringindo o acesso aos recursos do Amazon EC2. Para especificar permissões em nível de recurso para recursos especificados em um modelo de execução, é necessário incluir os recursos na declaração de ação`RunInstances`.

Neste exemplo, há duas declarações:

- A primeira declaração requer que os usuários executem instâncias em uma sub-rede específica (**subnet-1a2b3c4d**), usando um grupo de segurança (**sg-903004f88example**) específico e usando uma AMI (**ami-04d5cc9b88example**) específica. Ele também dá aos usuários acesso aos recursos especificados em um modelo de execução: interfaces de rede, pares de chaves e volumes.
- A segunda instrução permite que os usuários executem instâncias usando somente os tipos de instância **t2.micro** e,**t2.small** o que é possível fazer para controlar os custos.

  No entanto, observe que atualmente não há uma maneira eficaz de impedir completamente que os usuários que têm permissão para iniciar instâncias com um modelo de execução iniciem outros tipos de instância. Isso ocorre porque um tipo de instância especificado em um modelo de execução pode ser substituído para usar tipos de instância definidos usando a seleção de tipo de instância baseada em atributos.

Para obter uma lista completa das permissões em nível de recurso que você pode usar para controlar a configuração das instâncias que um usuário pode iniciar, consulte Ações, recursos e chaves de condição do Amazon EC2 na Referência de autorização do serviço.

Exemplo 2: exigência de tags e restrição do acesso a recursos para execuções de instâncias do Amazon EC2

A política de exemplo a seguir mostra como é possível usar condições em sua política baseada em identidade para controlar o acesso a recursos do Amazon EC2 com base em tags.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "InstanceTags",
            "Effect": "Allow",
            "Action": [
                "ec2:RunInstances"
            ],
            "Resource": [
                "arn:aws:ec2:us-east-1:555555555555:instance/*"
            ],
            "Condition": {
```

```
                "StringEquals": {
                    "aws:RequestTag/owner": "dev"
                }
            }
        },
        {
            "Sid": "InstanceBoundaries",
            "Effect": "Allow",
            "Action": [
                "ec2:RunInstances"
            ],
            "Resource": [
                "arn:aws:ec2:us-east-1::image/*",
                "arn:aws:ec2:us-east-1:555555555555:subnet/*",
                "arn:aws:ec2:us-east-1:555555555555:network-interface/*"
            ]
        },
        {
            "Sid": "InstanceResourceTags",
            "Effect": "Allow",
            "Action": [
                "ec2:RunInstances"
            ],
            "Resource": [
                "arn:aws:ec2:us-east-1:555555555555:key-pair/*",
                "arn:aws:ec2:us-east-1:555555555555:security-group/*",
                "arn:aws:ec2:us-east-1:555555555555:launch-template/*",
                "arn:aws:ec2:us-east-1:555555555555:volume/*"
            ],
            "Condition": {
                "StringEquals": {
                    "aws:ResourceTag/purpose": "testing"
                }
            }
        }
    ]
}
```

Neste exemplo, há três instruções:

- A primeira instrução permite que os usuários iniciem instâncias na região especificada somente se uma tag com `owner=dev` for usada na solicitação.

- A segunda instrução dá aos usuários acesso à AMI, à sub-rede e à interface de rede na região especificada.
- A terceira instrução permite que os usuários iniciem instâncias na região especificada com um par de chaves, um grupo de segurança, um modelo de execução e um volume existentes que tenham a tag `purpose=testing`.

Para obter mais informações, consulte Controle do acesso a AWS recursos usando tags no Guia do usuário do IAM.

Permissões necessárias para marcar instâncias e volumes

O exemplo a seguir permite que os usuários marquem instâncias e volumes na criação. Essa parte será necessária se houver tags especificadas no modelo de execução. Para obter mais informações, consulte Conceder permissão para marcar recursos durante a criação no Manual do usuário do Amazon EC2 para instâncias do Linux.

Permissões adicionais do modelo de execução

É necessário conceder permissões aos usuários do console para as ações `ec2:DescribeLaunchTemplates` e `ec2:DescribeLaunchTemplateVersions`. Sem essas permissões, os dados do modelo de execução não podem ser carregados no assistente do grupo do Auto Scaling, e os usuários não podem utilizar o assistente para iniciar instâncias usando um modelo de execução. É possível especificar essas ações adicionais no elemento `Action` de uma instrução de política do IAM.

Validação de permissões para **`ec2:RunInstances`** e **`iam:PassRole`**

Os usuários podem especificar qual versão de um modelo de execução seu grupo do Auto Scaling usa. Dependendo de suas permissões, essa pode ser uma versão numerada específica ou a versão `$Latest` ou `$Default` do modelo de execução. Se for o último, tome cuidado especial. Isso pode substituir as permissões `ec2:RunInstances` e `iam:PassRole` que você pretendia restringir.

Esta seção explica o cenário de uso da versão mais recente ou padrão do modelo de execução com um grupo do Auto Scaling.

Quando um usuário chama as APIs, `CreateAutoScalingGroup` `UpdateAutoScalingGroup` ou, `StartInstanceRefresh` o Amazon EC2 Auto Scaling verifica suas permissões em relação à versão do modelo de execução que é a versão mais recente ou padrão no momento antes de

prosseguir com a solicitação. Isso valida as permissões para ações a serem concluídas ao iniciar instâncias, como as ações `ec2:RunInstances` e `iam:PassRole`. Para fazer isso, emitimos uma chamada de RunInstancesdry run do Amazon EC2 para validar se o usuário tem as permissões necessárias para a ação, sem realmente fazer a solicitação. Quando uma resposta é retornada, ela é lida pelo Amazon EC2 Auto Scaling. Se as permissões do usuário não permitirem uma determinada ação, haverá falha na solicitação do Amazon EC2 Auto Scaling, que retornará um erro ao usuário contendo informações sobre a permissão ausente.

Depois que a verificação inicial e a solicitação forem concluídas, sempre que as instâncias forem executadas, o Amazon EC2 Auto Scaling as iniciará com a versão mais recente ou padrão, mesmo que ela tenha sido alterada, usando as permissões de seu perfil vinculado ao serviço. Isso significa que um usuário que esteja usando o modelo de execução pode atualizá-lo para transferir um perfil do IAM para uma instância, mesmo que não tenha a permissão `iam:PassRole`.

Use a chave de condição `autoscaling:LaunchTemplateVersionSpecified` se quiser limitar quem tem acesso à configuração de grupos para usar a versão `$Latest` ou `$Default`. Isso garante que o grupo Auto Scaling só aceite uma versão numerada específica quando um usuário chama e. `CreateAutoScalingGroup` `UpdateAutoScalingGroup` APIs Para ver um exemplo que mostra como adicionar essa chave de condição a uma política do IAM, consulte Exigir um modelo de execução e um número de versão.

Para grupos do Auto Scaling configurados para usar a versão `$Latest` ou `$Default` do modelo de execução, considere limitar quem pode criar e gerenciar versões do modelo de execução, incluindo a ação `ec2:ModifyLaunchTemplate` que permite ao usuário especificar a versão padrão do modelo de execução. Para obter mais informações, consulte Controlar permissões de versionamento no Guia do usuário do Amazon EC2 para instâncias do Linux.

Recursos relacionados

Para saber mais sobre permissões para visualizar, criar e excluir modelos de execução e versões de modelos de execução, consulte Controlar o acesso aos modelos de execução com permissões do IAM no Guia do usuário do Amazon EC2 para instâncias do Linux.

Para obter mais informações sobre as permissões em nível de recurso que você pode usar para controlar o acesso à chamada,`RunInstances` consulte Ações, recursos e chaves de condição do Amazon EC2 na Referência de autorização do serviço.

Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2

Aplicações executadas em instâncias do Amazon EC2 precisam de credenciais para acessar outros Serviços da AWS. Para fornecer essas credenciais de uma maneira segura, use um perfil do IAM. A função fornece permissões temporárias que a aplicação pode usar ao acessar outros recursos da AWS . As permissões da função determinam o que a aplicação tem permissão para fazer.

Para instâncias em um grupo do Auto Scaling, é necessário criar uma configuração de execução ou um modelo de execução e escolher um perfil de instância para associar às instâncias. Um perfil de instância é um contêiner para um perfil do IAM que permite ao Amazon EC2 passar o perfil do IAM para uma instância quando ela é iniciada. Primeiro, crie uma função do IAM que tenha todas as permissões necessárias para acessar os AWS recursos. Depois, crie o perfil da instância e atribua a função a ele.

> **Note**
>
> Como prática recomendada, é altamente recomendável que você crie a função para que ela tenha as permissões mínimas para outras Serviços da AWS que seu aplicativo exige.

Conteúdos

- Pré-requisitos
- Criar um modelo de execução
- Consulte também

Pré-requisitos

Crie o perfil do IAM que a aplicação em execução no Amazon EC2 pode assumir. Escolha as permissões apropriadas para que a aplicação que receber a função possa fazer as chamadas de API específicas necessárias.

Se você usa o console do IAM em vez do AWS CLI ou um dos AWS SDKs, o console cria um perfil de instância automaticamente e dá a ele o mesmo nome da função à qual ele corresponde.

Para criar um perfil do IAM (console)

1. Abra o console do IAM em https://console.aws.amazon.com/iam/.

2. No painel de navegação à esquerda, escolha Roles (Funções).
3. Selecione Criar perfil.
4. Em Select trusted entity (Selecionar entidade confiável), escolha AWS Service (Serviço).
5. Para seu caso de uso, escolha EC2 e escolha Next (Próximo).
6. Se possível, selecione a política a ser usada para a política de permissões ou escolha Create policy (Criar política) para abrir uma nova guia no navegador e criar uma nova política a partir do zero. Para obter mais informações, consulte Creating IAM policies (Criar políticas do IAM) no IAM User Guide (Guia do usuário do IAM). Depois de criar a política, feche essa guia e retorne à guia original. Marque a caixa de seleção ao lado das políticas de permissões que você deseja que o serviço tenha.
7. (Opcional) Defina um limite de permissões. Este é um recurso avançado que está disponível para funções de serviço. Para obter mais informações sobre limites de permissões, consulte Limites de permissões para identidades do IAM no Guia do usuário do IAM.
8. Escolha Próximo.
9. Na página Name, review, and create (Nomear, revisar e criar), em Role name (Nome do perfil), insira um nome de função para ajudar você a identificar a finalidade desse perfil. Esse nome deve ser exclusivo em sua Conta da AWS. Como outros AWS recursos podem fazer referência à função, você não pode editar o nome da função após sua criação.
10. Reveja a função e escolha Criar função.

permissões do IAM

Use uma política baseada em identidade do IAM para controlar o acesso ao novo perfil do IAM. A permissão `iam:PassRole` é necessária na identidade do IAM (usuário ou perfil) que cria ou atualiza um grupo do Auto Scaling usando um modelo de execução que especifica um perfil de instância.

O exemplo de política a seguir concede permissões para passar somente perfis do IAM cujo nome comece com **`qateam-`**.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
```

```
            "Action": "iam:PassRole",
            "Resource": "arn:aws:iam::111122223333:role/qateam-*",
            "Condition": {
                "StringEquals": {
                    "iam:PassedToService": [
                        "ec2.amazonaws.com",
                        "ec2.amazonaws.com.cn"
                    ]
                }
            }
        }
    ]
}
```

> **Important**
>
> Para obter informações sobre como o Amazon EC2 Auto Scaling valida permissões para a ação `iam:PassRole` de um grupo do Auto Scaling que usa um modelo de execução, consulte Validação de permissões para `ec2:RunInstances` e `iam:PassRole`.

Criar um modelo de execução

Ao criar o modelo de execução usando o Console de gerenciamento da AWS, na seção Detalhes avançados, selecione a função no perfil da instância do IAM. Para obter mais informações, consulte Criar um modelo de execução usando configurações avançadas.

Ao criar o modelo de execução usando o create-launch-templatecomando do AWS CLI, especifique o nome do perfil da instância da sua função do IAM, conforme mostrado no exemplo a seguir.

```
aws ec2 create-launch-template --launch-template-name my-lt-with-instance-profile --
version-description version1 \
--launch-template-data
 '{"ImageId":"ami-04d5cc9b88example","InstanceType":"t2.micro","IamInstanceProfile":
{"Name":"my-instance-profile"}}'
```

Consulte também

Para obter mais informações para começar a aprender e usar perfis do IAM para Amazon EC2, consulte:

- Políticas do IAM para o Amazon EC2 no Guia do usuário do Amazon EC2.
- Uso de perfis de instância e Uso de perfis do IAM para conceder permissões a aplicações em execução nas instâncias do Amazon EC2 no Manual do usuário do IAM

Validação de compatibilidade do Amazon EC2 Auto Scaling

Para saber se um AWS service (Serviço da AWS) está dentro do escopo de programas de conformidade específicos, consulte Serviços da AWS Escopo por Programa de Conformidade Serviços da AWS e escolha o programa de conformidade em que você está interessado. Para obter informações gerais, consulte Programas de AWS conformidade Programas AWS de .

Você pode baixar relatórios de auditoria de terceiros usando AWS Artifact. Para obter mais informações, consulte Baixar relatórios em AWS Artifact .

Sua responsabilidade de conformidade ao usar Serviços da AWS é determinada pela confidencialidade de seus dados, pelos objetivos de conformidade de sua empresa e pelas leis e regulamentos aplicáveis. Para obter mais informações sobre sua responsabilidade de conformidade ao usar Serviços da AWS, consulte a documentação AWS de segurança.

Conformidade do PCI DSS

O Amazon EC2 Auto Scaling é compatível com o processamento, o armazenamento e a transmissão de dados de cartão de crédito por comerciantes ou provedores de serviços e foi validado como compatível com o Data Security Standard (DSS, Padrão de segurança de dados) da Payment Card Industry (PCI). Para obter mais informações sobre o PCI DSS, incluindo como solicitar uma cópia do PCI AWS Compliance Package, consulte PCI DSS Nível 1.

Para obter informações sobre como alcançar a conformidade com o PCI DSS para suas AWS cargas de trabalho, consulte o seguinte guia de conformidade:

- Padrão de Segurança de Dados do Setor de Cartões de Pagamento (PCI DSS) 3.2.1 ativado AWS

Amazon EC2 Auto Scaling e endpoints da VPC da interface

É possível melhorar o procedimento de segurança da sua VPC configurando o Amazon EC2 Auto Scaling para usar um endpoint da VPC de interface. Os endpoints de interface são alimentados por AWS PrivateLink uma tecnologia que permite que você acesse de forma privada as APIs do Amazon

EC2 Auto Scaling restringindo todo o tráfego de rede entre sua VPC e o Amazon EC2 Auto Scaling à rede. AWS Com endpoints de interface, também não são necessários um gateway da Internet, um dispositivo NAT nem um gateway privado virtual.

Não é necessário configurar AWS PrivateLink, mas é recomendado. Para obter mais informações sobre AWS PrivateLink endpoints de VPC, consulte O que é? AWS PrivateLink no AWS PrivateLink Guia.

Tópicos

- Criar um VPC endpoint de interface
- Criar uma política de endpoint da VPC

Criar um VPC endpoint de interface

Crie um endpoint para o Amazon EC2 Auto Scaling usando o seguinte nome de serviço:

```
com.amazonaws.region.autoscaling
```

Para obter mais informações, consulte Acessar um AWS serviço usando uma interface VPC endpoint no Guia.AWS PrivateLink

Você não precisa alterar nenhuma configuração do Amazon EC2 Auto Scaling. O Amazon EC2 Auto Scaling AWS chama outros serviços usando endpoints de serviço ou endpoints VPC de interface privada, os que estiverem em uso.

Criar uma política de endpoint da VPC

É possível associar uma política ao seu endpoint da VPC para controlar o acesso à API do Amazon EC2 Auto Scaling. A política especifica:

- O principal que pode executar ações.
- As ações que podem ser executadas.
- O recurso no qual as ações podem ser executadas.

O exemplo a seguir mostra uma política de VPC endpoint que nega a todos permissão para excluir uma política de escalabilidade por meio do endpoint. O exemplo de política também concede a todos permissão para executar todas as outras ações.

```
{
   "Statement": [
        {
            "Action": "*",
            "Effect": "Allow",
            "Resource": "*",
            "Principal": "*"
        },
        {
            "Action": "autoscaling:DeleteScalingPolicy",
            "Effect": "Deny",
            "Resource": "*",
            "Principal": "*"
        }
    ]
}
```

Para obter mais informações, consulte Controlar o acesso aos endpoints da VPC usando políticas de endpoint no Guia AWS PrivateLink .

Usando esse serviço com um AWS SDK

AWS kits de desenvolvimento de software (SDKs) estão disponíveis para muitas linguagens de programação populares. Cada SDK fornece uma API, exemplos de código e documentação que permitem que os desenvolvedores criem facilmente aplicações em seu idioma de preferência.

Documentação do SDK	Exemplos de código
AWS SDK para C++	AWS SDK para C++ exemplos de código
AWS CLI	AWS CLI exemplos de código
AWS SDK para Go	AWS SDK para Go exemplos de código
AWS SDK para Java	AWS SDK para Java exemplos de código
AWS SDK para JavaScript	AWS SDK para JavaScript exemplos de código
AWS SDK para Kotlin	AWS SDK para Kotlin exemplos de código
AWS SDK para .NET	AWS SDK para .NET exemplos de código
AWS SDK para PHP	AWS SDK para PHP exemplos de código
Ferramentas da AWS para PowerShell	Ferramentas da AWS para PowerShell exemplos de código
AWS SDK para Python (Boto3)	AWS SDK para Python (Boto3) exemplos de código
AWS SDK para Ruby	AWS SDK para Ruby exemplos de código
AWS SDK para Rust	AWS SDK para Rust exemplos de código
AWS SDK para SAP ABAP	AWS SDK para SAP ABAP exemplos de código
AWS SDK para Swift	AWS SDK para Swift exemplos de código

Para obter exemplos específicos deste serviço, consulte Exemplos de código para Auto Scaling usando AWS SDKs.

> **Exemplo de disponibilidade**
>
> Não consegue encontrar o que precisa? Solicite um exemplo de código usando o link Fornecer feedback na parte inferior desta página.

Trabalhando com as APIs do Amazon EC2 Auto Scaling

A API do Amazon EC2 Auto Scaling é a interface em nível de protocolo para o Amazon EC2 Auto Scaling. Ao usar a API, você deve formatar cada solicitação HTTPS corretamente e adicionar uma assinatura digital válida a cada solicitação. Para obter mais informações, consulte a referência da API Amazon EC2 Auto Scaling.

A API do Amazon EC2 Auto Scaling inclui as seguintes ações de API:

Limites da conta

- DescribeAccountLimits

Grupos de Auto Scaling

- AttachLoadBalancers
- AttachLoadBalancerTargetGroups
- CreateAutoScalingGroup
- DeleteAutoScalingGroup
- DescribeAutoScalingGroups
- DescribeLoadBalancers
- DescribeLoadBalancerTargetGroups
- DetachLoadBalancerTargetGroups
- DetachLoadBalancers
- UpdateAutoScalingGroup

Instâncias de Auto Scaling

- AttachInstances
- DescribeAutoScalingInstances
- DetachInstances
- SetInstanceHealth
- SetInstanceProtection

- TerminateInstanceInAutoScalingGroup

Atualizações de instâncias

- CancelInstanceRefresh
- DescribeInstanceRefreshes
- RollbackInstanceRefresh
- StartInstanceRefresh

Configurações de execução

- CreateLaunchConfiguration
- DeleteLaunchConfiguration
- DescribeLaunchConfigurations

Ganchos do ciclo de vida

- CompleteLifecycleAction
- DeleteLifecycleHook
- DescribeLifecycleHooks
- DescribeLifecycleHookTypes
- PutLifecycleHook
- RecordLifecycleActionHeartbeat

Monitoramento

- DeleteNotificationConfiguration
- DescribeAutoScalingNotificationTypes
- DescribeMetricCollectionTypes
- DescribeNotificationConfigurations
- DisableMetricsCollection
- EnableMetricsCollection
- PutNotificationConfiguration

Escalabilidade

- DeletePolicy
- DescribeAdjustmentTypes
- DescribePolicies
- DescribeScalingActivities
- DescribeScalingProcessTypes
- DescribeTerminationPolicyTypes
- ExecutePolicy
- PutScalingPolicy
- ResumeProcesses
- SetDesiredCapacity
- SuspendProcesses

Escalabilidade programada

- BatchDeleteScheduledAction
- BatchPutScheduledUpdateGroupAction
- DeleteScheduledAction
- DescribeScheduledActions
- PutScheduledUpdateGroupAction

Estado de espera

- EnterStandby
- ExitStandby

Tags

- CreateOrUpdateTags
- DeleteTags
- DescribeTags

Piscinas quentes

- PutWarmPool
- DescribeWarmPool
- DeleteWarmPool

Exemplos de código para Auto Scaling usando AWS SDKs

Os exemplos de código a seguir mostram como usar o Auto Scaling com um kit de desenvolvimento AWS de software (SDK).

As noções básicas são exemplos de código que mostram como realizar as operações essenciais em um serviço.

Ações são trechos de código de programas maiores e devem ser executadas em contexto. Embora as ações mostrem como chamar perfis de serviço individuais, você pode ver as ações no contexto em seus cenários relacionados.

Cenários são exemplos de código que mostram como realizar tarefas específicas chamando várias funções dentro de um serviço ou combinadas com outros Serviços da AWS.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Exemplos de código

- Exemplos básicos de Auto Scaling usando AWS SDKs
  - Olá, Auto Scaling
  - Aprenda as noções básicas do Auto Scaling com um SDK AWS
  - Ações para Auto Scaling usando AWS SDKs
    - Usar AttachInstances com uma CLI
    - Use AttachLoadBalancerTargetGroups com um AWS SDK ou CLI
    - Usar AttachLoadBalancers com uma CLI
    - Usar CompleteLifecycleAction com uma CLI
    - Use CreateAutoScalingGroup com um AWS SDK ou CLI
    - Usar CreateLaunchConfiguration com uma CLI
    - Usar CreateOrUpdateTags com uma CLI
    - Use DeleteAutoScalingGroup com um AWS SDK ou CLI
    - Usar DeleteLaunchConfiguration com uma CLI
    - Usar DeleteLifecycleHook com uma CLI
    - Usar DeleteNotificationConfiguration com uma CLI

- Usar DeletePolicy com uma CLI
- Usar DeleteScheduledAction com uma CLI
- Usar DeleteTags com uma CLI
- Usar DescribeAccountLimits com uma CLI
- Usar DescribeAdjustmentTypes com uma CLI
- Use DescribeAutoScalingGroups com um AWS SDK ou CLI
- Use DescribeAutoScalingInstances com um AWS SDK ou CLI
- Usar DescribeAutoScalingNotificationTypes com uma CLI
- Usar DescribeLaunchConfigurations com uma CLI
- Usar DescribeLifecycleHookTypes com uma CLI
- Usar DescribeLifecycleHooks com uma CLI
- Usar DescribeLoadBalancers com uma CLI
- Usar DescribeMetricCollectionTypes com uma CLI
- Usar DescribeNotificationConfigurations com uma CLI
- Usar DescribePolicies com uma CLI
- Use DescribeScalingActivities com um AWS SDK ou CLI
- Usar DescribeScalingProcessTypes com uma CLI
- Usar DescribeScheduledActions com uma CLI
- Usar DescribeTags com uma CLI
- Usar DescribeTerminationPolicyTypes com uma CLI
- Usar DetachInstances com uma CLI
- Usar DetachLoadBalancers com uma CLI
- Use DisableMetricsCollection com um AWS SDK ou CLI
- Use EnableMetricsCollection com um AWS SDK ou CLI
- Usar EnterStandby com uma CLI
- Usar ExecutePolicy com uma CLI
- Usar ExitStandby com uma CLI
- Usar PutLifecycleHook com uma CLI
- Usar PutNotificationConfiguration com uma CLI
- Usar PutScalingPolicy com uma CLI

    - Usar PutScheduledUpdateGroupAction com uma CLI
    - Usar RecordLifecycleActionHeartbeat com uma CLI
    - Usar ResumeProcesses com uma CLI
    - Use SetDesiredCapacity com um AWS SDK ou CLI
    - Usar SetInstanceHealth com uma CLI
    - Usar SetInstanceProtection com uma CLI
    - Usar SuspendProcesses com uma CLI
    - Use TerminateInstanceInAutoScalingGroup com um AWS SDK ou CLI
    - Use UpdateAutoScalingGroup com um AWS SDK ou CLI
- Cenários para o Auto Scaling usando AWS SDKs
    - Crie e gerencie um serviço resiliente usando um SDK AWS

Exemplos básicos de Auto Scaling usando AWS SDKs

Os exemplos de código a seguir mostram como usar os conceitos básicos do Amazon EC2 Auto Scaling com. AWS SDKs

Exemplos

- Olá, Auto Scaling
- Aprenda as noções básicas do Auto Scaling com um SDK AWS
- Ações para Auto Scaling usando AWS SDKs
    - Usar AttachInstances com uma CLI
    - Use AttachLoadBalancerTargetGroups com um AWS SDK ou CLI
    - Usar AttachLoadBalancers com uma CLI
    - Usar CompleteLifecycleAction com uma CLI
    - Use CreateAutoScalingGroup com um AWS SDK ou CLI
    - Usar CreateLaunchConfiguration com uma CLI
    - Usar CreateOrUpdateTags com uma CLI
    - Use DeleteAutoScalingGroup com um AWS SDK ou CLI
    - Usar DeleteLaunchConfiguration com uma CLI
    - Usar DeleteLifecycleHook com uma CLI

- Usar DeleteNotificationConfiguration com uma CLI
- Usar DeletePolicy com uma CLI
- Usar DeleteScheduledAction com uma CLI
- Usar DeleteTags com uma CLI
- Usar DescribeAccountLimits com uma CLI
- Usar DescribeAdjustmentTypes com uma CLI
- Use DescribeAutoScalingGroups com um AWS SDK ou CLI
- Use DescribeAutoScalingInstances com um AWS SDK ou CLI
- Usar DescribeAutoScalingNotificationTypes com uma CLI
- Usar DescribeLaunchConfigurations com uma CLI
- Usar DescribeLifecycleHookTypes com uma CLI
- Usar DescribeLifecycleHooks com uma CLI
- Usar DescribeLoadBalancers com uma CLI
- Usar DescribeMetricCollectionTypes com uma CLI
- Usar DescribeNotificationConfigurations com uma CLI
- Usar DescribePolicies com uma CLI
- Use DescribeScalingActivities com um AWS SDK ou CLI
- Usar DescribeScalingProcessTypes com uma CLI
- Usar DescribeScheduledActions com uma CLI
- Usar DescribeTags com uma CLI
- Usar DescribeTerminationPolicyTypes com uma CLI
- Usar DetachInstances com uma CLI
- Usar DetachLoadBalancers com uma CLI
- Use DisableMetricsCollection com um AWS SDK ou CLI
- Use EnableMetricsCollection com um AWS SDK ou CLI
- Usar EnterStandby com uma CLI
- Usar ExecutePolicy com uma CLI
- Usar ExitStandby com uma CLI
- Usar PutLifecycleHook com uma CLI
- Usar PutNotificationConfiguration com uma CLI

- Usar PutScalingPolicy com uma CLI
- Usar PutScheduledUpdateGroupAction com uma CLI
- Usar RecordLifecycleActionHeartbeat com uma CLI
- Usar ResumeProcesses com uma CLI
- Use SetDesiredCapacity com um AWS SDK ou CLI
- Usar SetInstanceHealth com uma CLI
- Usar SetInstanceProtection com uma CLI
- Usar SuspendProcesses com uma CLI
- Use TerminateInstanceInAutoScalingGroup com um AWS SDK ou CLI
- Use UpdateAutoScalingGroup com um AWS SDK ou CLI

Olá, Auto Scaling

Os exemplos de código a seguir mostram como começar a usar o Auto Scaling.

.NET

SDK para .NET (v4)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
namespace AutoScalingActions;

using Amazon.AutoScaling;

public class HelloAutoScaling
{
    /// <summary>
    /// Hello Amazon EC2 Auto Scaling. List EC2 Auto Scaling groups.
    /// </summary>
    /// <param name="args"></param>
    /// <returns>Async Task.</returns>
```

```
    static async Task Main(string[] args)
    {
        var client = new AmazonAutoScalingClient();

        Console.WriteLine("Welcome to Amazon EC2 Auto Scaling.");
        Console.WriteLine("Let's get a description of your Auto Scaling
groups.");

        var response = await client.DescribeAutoScalingGroupsAsync();

        if (response.AutoScalingGroups == null ||
response.AutoScalingGroups.Count == 0)
        {
            Console.WriteLine("Sorry, you don't have any Amazon EC2 Auto Scaling
groups.");
            return;
        }
        response.AutoScalingGroups.ForEach(autoScalingGroup =>
        {

Console.WriteLine($"{autoScalingGroup.AutoScalingGroupName}\t{autoScalingGroup.Availabil
        });

    }
}
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK para .NET da API.

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

Código para o CMake arquivo CMake Lists.txt.

```
# Set the minimum required version of CMake for this project.
cmake_minimum_required(VERSION 3.13)

# Set the AWS service components used by this project.
set(SERVICE_COMPONENTS autoscaling)

# Set this project's name.
project("hello_autoscaling")

# Set the C++ standard to use to build this target.
# At least C++ 11 is required for the AWS SDK for C++.
set(CMAKE_CXX_STANDARD 11)

# Use the MSVC variable to determine if this is a Windows build.
set(WINDOWS_BUILD ${MSVC})

if (WINDOWS_BUILD) # Set the location where CMake can find the installed
 libraries for the AWS SDK.
    string(REPLACE ";" "/aws-cpp-sdk-all;" SYSTEM_MODULE_PATH
 "${CMAKE_SYSTEM_PREFIX_PATH}/aws-cpp-sdk-all")
    list(APPEND CMAKE_PREFIX_PATH ${SYSTEM_MODULE_PATH})
endif ()

# Find the AWS SDK for C++ package.
find_package(AWSSDK REQUIRED COMPONENTS ${SERVICE_COMPONENTS})

if (WINDOWS_BUILD AND AWSSDK_INSTALL_AS_SHARED_LIBS)
     # Copy relevant AWS SDK for C++ libraries into the current binary directory
 for running and debugging.

     # set(BIN_SUB_DIR "/Debug") # If you are building from the command line, you
 may need to uncomment this
                                     # and set the proper subdirectory to the
 executables' location.

     AWSSDK_CPY_DYN_LIBS(SERVICE_COMPONENTS ""
 ${CMAKE_CURRENT_BINARY_DIR}${BIN_SUB_DIR})
endif ()

add_executable(${PROJECT_NAME}
        hello_autoscaling.cpp)

target_link_libraries(${PROJECT_NAME}
```

```
        ${AWSSDK_LINK_LIBRARIES})
```

Código para o arquivo de origem hello_autoscalling.cpp.

```
#include <aws/core/Aws.h>
#include <aws/autoscaling/AutoScalingClient.h>
#include <aws/autoscaling/model/DescribeAutoScalingGroupsRequest.h>
#include <iostream>

/*
 *  A "Hello Autoscaling" starter application which initializes an Amazon EC2
 Auto Scaling client and describes the
 *  Amazon EC2 Auto Scaling groups.
 *
 *  main function
 *
 *  Usage: 'hello_autoscaling'
 *
 */

int main(int argc, char **argv) {
    Aws::SDKOptions options;
    // Optionally change the log level for debugging.
//   options.loggingOptions.logLevel = Utils::Logging::LogLevel::Debug;
    Aws::InitAPI(options); // Should only be called once.
    int result = 0;
    {
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

        Aws::AutoScaling::AutoScalingClient autoscalingClient(clientConfig);

        std::vector<Aws::String> groupNames;
        Aws::String nextToken; // Used for pagination.

        do {

            Aws::AutoScaling::Model::DescribeAutoScalingGroupsRequest request;
            if (!nextToken.empty()) {
                request.SetNextToken(nextToken);
            }
```

```
            Aws::AutoScaling::Model::DescribeAutoScalingGroupsOutcome outcome =
                    autoscalingClient.DescribeAutoScalingGroups(request);

            if (outcome.IsSuccess()) {
                const Aws::Vector<Aws::AutoScaling::Model::AutoScalingGroup>
 &autoScalingGroups =
                        outcome.GetResult().GetAutoScalingGroups();
                for (auto &group: autoScalingGroups) {
                    groupNames.push_back(group.GetAutoScalingGroupName());
                }
                nextToken = outcome.GetResult().GetNextToken();
            } else {
                std::cerr << "Error with AutoScaling::DescribeAutoScalingGroups.
 "
                          << outcome.GetError().GetMessage()
                          << std::endl;
                result = 1;
                break;
            }
        } while (!nextToken.empty());

        std::cout << "Found " << groupNames.size() << " AutoScaling groups." <<
 std::endl;
        for (auto &groupName: groupNames) {
            std::cout << "AutoScaling group: " << groupName << std::endl;
        }

    }


    Aws::ShutdownAPI(options); // Should only be called once.
    return result;
}
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK para C++ da API.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.autoscaling.AutoScalingClient;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingGroup;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsResponse;
import java.util.List;

/**
 * Before running this SDK for Java (v2) code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class DescribeAutoScalingGroups {
    public static void main(String[] args) throws InterruptedException {
        AutoScalingClient autoScalingClient = AutoScalingClient.builder()
                .region(Region.US_EAST_1)
                .build();

        describeGroups(autoScalingClient);
    }

    public static void describeGroups(AutoScalingClient autoScalingClient) {
        DescribeAutoScalingGroupsResponse response =
 autoScalingClient.describeAutoScalingGroups();
        List<AutoScalingGroup> groups = response.autoScalingGroups();
        groups.forEach(group -> {
            System.out.println("Group Name: " + group.autoScalingGroupName());
            System.out.println("Group ARN: " + group.autoScalingGroupARN());
```

```
        });
    }
}
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK for Java 2.x da API.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function helloService()
    {
        $autoScalingClient = new AutoScalingClient([
            'region' => 'us-west-2',
            'version' => 'latest',
            'profile' => 'default',
        ]);

        $groups = $autoScalingClient->describeAutoScalingGroups([]);
        var_dump($groups);
    }
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK para PHP da API.

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
import boto3


def hello_autoscaling(autoscaling_client):
    """
    Use the AWS SDK for Python (Boto3) to create an Amazon EC2 Auto Scaling
 client and list
    some of the Auto Scaling groups in your account.
    This example uses the default settings specified in your shared credentials
    and config files.

    :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client object.
    """
    print(
        "Hello, Amazon EC2 Auto Scaling! Let's list up to ten of you Auto Scaling
 groups:"
    )
    response = autoscaling_client.describe_auto_scaling_groups()
    groups = response.get("AutoScalingGroups", [])
    if groups:
        for group in groups:
            print(f"\t{group['AutoScalingGroupName']}:
 {group['AvailabilityZones']}")
    else:
        print("There are no Auto Scaling groups in your account.")


if __name__ == "__main__":
    hello_autoscaling(boto3.client("autoscaling"))
```

- Para obter detalhes da API, consulte a DescribeAutoScalingGroupsReferência da API AWS SDK for Python (Boto3).

Ruby

SDK para Ruby

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
require 'aws-sdk-autoscaling'
require 'logger'

# AutoScalingManager is a class responsible for managing AWS Auto Scaling
 operations
# such as listing all Auto Scaling groups in the current AWS account.
class AutoScalingManager
  def initialize(client)
    @client = client
    @logger = Logger.new($stdout)
  end

  # Gets and prints a list of Auto Scaling groups for the account.
  def list_auto_scaling_groups
    paginator = @client.describe_auto_scaling_groups
    auto_scaling_groups = []
    paginator.each_page do |page|
      auto_scaling_groups.concat(page.auto_scaling_groups)
    end

    if auto_scaling_groups.empty?
      @logger.info('No Auto Scaling groups found for this account.')
    else
      auto_scaling_groups.each do |group|
        @logger.info("Auto Scaling group name: #{group.auto_scaling_group_name}")
        @logger.info("  Group ARN:             #{group.auto_scaling_group_arn}")
```

```
        @logger.info("  Min/max/desired:       #{group.min_size}/
#{group.max_size}/#{group.desired_capacity}")
        @logger.info("\n")
      end
    end
  end
end

if $PROGRAM_NAME == __FILE__
  autoscaling_client = Aws::AutoScaling::Client.new
  manager = AutoScalingManager.new(autoscaling_client)
  manager.list_auto_scaling_groups
end
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK para Ruby da API.

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
async fn list_groups(client: &Client) -> Result<(), Error> {
    let resp = client.describe_auto_scaling_groups().send().await?;

    println!("Groups:");

    let groups = resp.auto_scaling_groups();

    for group in groups {
        println!(
            "Name:  {}",
            group.auto_scaling_group_name().unwrap_or("Unknown")
        );
```

```
        println!(
            "Arn:   {}",
            group.auto_scaling_group_arn().unwrap_or("unknown"),
        );
        println!("Zones: {:?}", group.availability_zones(),);
        println!();
    }

    println!("Found {} group(s)", groups.len());

    Ok(())
}
```

- Para obter detalhes da API, consulte a DescribeAutoScalingGroupsreferência da API AWS SDK for Rust.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Aprenda as noções básicas do Auto Scaling com um SDK AWS

Os exemplos de código a seguir mostram como:

- Crie um grupo do Amazon EC2 Auto Scaling com um modelo de inicialização e zonas de disponibilidade e obtenha informações sobre instâncias em execução.
- Ative a coleta de CloudWatch métricas da Amazon.
- Atualizar a capacidade desejada do grupo e aguardar a inicialização de uma instância.
- Encerrar uma instância no grupo.
- Listar as atividades de ajuste de escala que ocorrem em resposta às solicitações do usuário e às mudanças de capacidade.
- Obtenha estatísticas de CloudWatch métricas e, em seguida, limpe os recursos.

.NET

SDK para .NET (v4)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
global using Amazon.AutoScaling;
global using Amazon.AutoScaling.Model;
global using Amazon.CloudWatch;
global using AutoScalingActions;
global using Microsoft.Extensions.DependencyInjection;
global using Microsoft.Extensions.Hosting;
global using Microsoft.Extensions.Logging;
global using Microsoft.Extensions.Logging.Console;
global using Microsoft.Extensions.Logging.Debug;



using Amazon.EC2;
using Microsoft.Extensions.Configuration;
using Host = Microsoft.Extensions.Hosting.Host;

namespace AutoScalingBasics;

public class AutoScalingBasics
{

    static async Task Main(string[] args)
    {
        // Set up dependency injection for Amazon EC2 Auto Scaling, Amazon
        // CloudWatch, and Amazon EC2.
        using var host = Host.CreateDefaultBuilder(args)
            .ConfigureLogging(logging =>
                logging.AddFilter("System", LogLevel.Debug)
                    .AddFilter<DebugLoggerProvider>("Microsoft",
 LogLevel.Information)
                    .AddFilter<ConsoleLoggerProvider>("Microsoft",
 LogLevel.Trace))
```

```
            .ConfigureServices((_, services) =>
            services.AddAWSService<IAmazonAutoScaling>()
                .AddAWSService<IAmazonCloudWatch>()
                .AddAWSService<IAmazonEC2>()
                .AddTransient<AutoScalingWrapper>()
                .AddTransient<CloudWatchWrapper>()
                .AddTransient<EC2Wrapper>()
                .AddTransient<UIWrapper>()
            )
            .Build();


        var autoScalingWrapper =
host.Services.GetRequiredService<AutoScalingWrapper>();
        var cloudWatchWrapper =
host.Services.GetRequiredService<CloudWatchWrapper>();
        var ec2Wrapper = host.Services.GetRequiredService<EC2Wrapper>();
        var uiWrapper = host.Services.GetRequiredService<UIWrapper>();

        var configuration = new ConfigurationBuilder()
            .SetBasePath(Directory.GetCurrentDirectory())
            .AddJsonFile("settings.json") // Load test settings from .json file.
            .AddJsonFile("settings.local.json",
                true) // Optionally load local settings.
            .Build();

        var imageId = configuration["ImageId"];
        var instanceType = configuration["InstanceType"];
        var launchTemplateName = configuration["LaunchTemplateName"];

        launchTemplateName += Guid.NewGuid().ToString();

        // The name of the Auto Scaling group.
        var groupName = configuration["GroupName"];

        uiWrapper.DisplayTitle("Auto Scaling Basics");
        uiWrapper.DisplayAutoScalingBasicsDescription();

        // Create the launch template and save the template Id to use when
deleting the
        // launch template at the end of the application.
        var launchTemplateId = await
ec2Wrapper.CreateLaunchTemplateAsync(imageId!, instanceType!,
launchTemplateName);
```

```
        // Confirm that the template was created by asking for a description of
it.
        await ec2Wrapper.DescribeLaunchTemplateAsync(launchTemplateName);

        uiWrapper.PressEnter();

        var availabilityZones = await ec2Wrapper.ListAvailabilityZonesAsync();

        Console.WriteLine($"Creating an Auto Scaling group named {groupName}.");
        await autoScalingWrapper.CreateAutoScalingGroupAsync(
            groupName!,
            launchTemplateName,
            availabilityZones[0].ZoneName);

        // Keep checking the details of the new group until its lifecycle state
        // is "InService".
        Console.WriteLine($"Waiting for the Auto Scaling group to be active.");

        List<AutoScalingInstanceDetails> instanceDetails;

        do
        {
            instanceDetails = await
autoScalingWrapper.DescribeAutoScalingInstancesAsync(groupName!);
        }
        while (instanceDetails.Count <= 0);

        Console.WriteLine($"Auto scaling group {groupName} successfully
created.");
        Console.WriteLine($"{instanceDetails.Count} instances were created for
the group.");

        // Display the details of the Auto Scaling group.
        instanceDetails.ForEach(detail =>
        {
            Console.WriteLine($"Group name: {detail.AutoScalingGroupName}");
        });

        uiWrapper.PressEnter();

        uiWrapper.DisplayTitle("Metrics collection");
        Console.WriteLine($"Enable metrics collection for {groupName}");
        await autoScalingWrapper.EnableMetricsCollectionAsync(groupName!);
```

```
        // Show the metrics that are collected for the group.

        // Update the maximum size of the group to three instances.
        Console.WriteLine("--- Update the Auto Scaling group to increase max size
to 3 ---");
        int maxSize = 3;
        await autoScalingWrapper.UpdateAutoScalingGroupAsync(groupName!,
launchTemplateName, maxSize);

        Console.WriteLine("--- Describe all Auto Scaling groups to show the
current state of the group ---");
        var groups = await
autoScalingWrapper.DescribeAutoScalingGroupsAsync(groupName!);

        uiWrapper.DisplayGroupDetails(groups!);

        uiWrapper.PressEnter();

        uiWrapper.DisplayTitle("Describe account limits");
        await autoScalingWrapper.DescribeAccountLimitsAsync();

        uiWrapper.WaitABit(60, "Waiting for the resources to be ready.");

        uiWrapper.DisplayTitle("Set desired capacity");
        int desiredCapacity = 2;
        await autoScalingWrapper.SetDesiredCapacityAsync(groupName!,
desiredCapacity);

        Console.WriteLine("Get the two instance Id values");

        // Empty the group before getting the details again.
        groups.Clear();
        groups = await
autoScalingWrapper.DescribeAutoScalingGroupsAsync(groupName!);
        if (groups.Any())
        {
            foreach (AutoScalingGroup group in groups)
            {
                Console.WriteLine($"The group name is
{group.AutoScalingGroupName}");
                Console.WriteLine($"The group ARN is
{group.AutoScalingGroupARN}");
                var instances = group.Instances;
```

```
                foreach (Amazon.AutoScaling.Model.Instance instance in instances)
                {
                    Console.WriteLine($"The instance id is
{instance.InstanceId}");
                    Console.WriteLine($"The lifecycle state is
{instance.LifecycleState}");
                }
            }
        }

        uiWrapper.DisplayTitle("Scaling Activities");
        Console.WriteLine("Let's list the scaling activities that have occurred
for the group.");
        var activities = await
autoScalingWrapper.DescribeScalingActivitiesAsync(groupName!);
        if (activities.Any())
        {
            activities.ForEach(activity =>
            {
                Console.WriteLine($"The activity Id is {activity.ActivityId}");
                Console.WriteLine($"The activity details are
{activity.Details}");
            });
        }

        // Display the Amazon CloudWatch metrics that have been collected.
        var metrics = await
cloudWatchWrapper.GetCloudWatchMetricsAsync(groupName!);
        if (metrics.Any())
        {
            Console.WriteLine($"Metrics collected for {groupName}:");
            metrics.ForEach(metric =>
            {
                Console.Write($"Metric name: {metric.MetricName}\t");
                Console.WriteLine($"Namespace: {metric.Namespace}");
            });
        }

        var dataPoints = await
cloudWatchWrapper.GetMetricStatisticsAsync(groupName!);
        if (dataPoints.Any())
        {
            Console.WriteLine("Details for the metrics collected:");
            dataPoints.ForEach(detail => { Console.WriteLine(detail); });
```

```
        }

        // Disable metrics collection.
        Console.WriteLine("Disabling the collection of metrics for
{groupName}.");
        var success = await
autoScalingWrapper.DisableMetricsCollectionAsync(groupName!);

        if (success)
        {
            Console.WriteLine($"Successfully stopped metrics collection for
{groupName}.");
        }
        else
        {
            Console.WriteLine($"Could not stop metrics collection for
{groupName}.");
        }

        // Terminate all instances in the group.
        uiWrapper.DisplayTitle("Terminating Auto Scaling instances");
        Console.WriteLine("Now terminating all instances in the Auto Scaling
group.");

        if (groups is not null)
        {
            groups.ForEach(group =>
            {
                // Only delete instances in the AutoScaling group we created.
                if (group.AutoScalingGroupName == groupName)
                {
                    group.Instances.ForEach(async instance =>
                    {
                        await
autoScalingWrapper.TerminateInstanceInAutoScalingGroupAsync(instance.InstanceId);
                    });
                }
            });
        }

        // After all instances are terminated, delete the group.
        uiWrapper.DisplayTitle("Clean up resources");
        Console.WriteLine("Deleting the Auto Scaling group.");
        await autoScalingWrapper.DeleteAutoScalingGroupAsync(groupName!);
```

```
        // Delete the launch template.
        var deletedLaunchTemplateName = await
 ec2Wrapper.DeleteLaunchTemplateAsync(launchTemplateId);

        if (deletedLaunchTemplateName == launchTemplateName)
        {
            Console.WriteLine("Successfully deleted the launch template.");
        }

        Console.WriteLine("The demo is now concluded.");
    }
}


namespace AutoScalingBasics;

/// <summary>
/// A class to provide user interface methods for the EC2 AutoScaling Basics
/// scenario.
/// </summary>
public class UIWrapper
{
    public readonly string SepBar = new('-', Console.WindowWidth);

    /// <summary>
    /// Describe the steps in the EC2 AutoScaling Basics scenario.
    /// </summary>
    public void DisplayAutoScalingBasicsDescription()
    {
        Console.WriteLine("This code example performs the following
 operations:");
        Console.WriteLine(" 1. Creates an Amazon EC2 launch template.");
        Console.WriteLine(" 2. Creates an Auto Scaling group.");
        Console.WriteLine(" 3. Shows the details of the new Auto Scaling group");
        Console.WriteLine("    to show that only one instance was created.");
        Console.WriteLine(" 4. Enables metrics collection.");
        Console.WriteLine(" 5. Updates the Auto Scaling group to increase the");
        Console.WriteLine("    capacity to three.");
        Console.WriteLine(" 6. Describes Auto Scaling groups again to show the");
        Console.WriteLine("    current state of the group.");
        Console.WriteLine(" 7. Changes the desired capacity of the Auto
 Scaling");
        Console.WriteLine("    group to use an additional instance.");
```

```
        Console.WriteLine(" 8. Shows that there are now instances in the
 group.");
        Console.WriteLine(" 9. Lists the scaling activities that have occurred
 for the group.");
        Console.WriteLine("10. Displays the Amazon CloudWatch metrics that
 have");
        Console.WriteLine("    been collected.");
        Console.WriteLine("11. Disables metrics collection.");
        Console.WriteLine("12. Terminates all instances in the Auto Scaling
 group.");
        Console.WriteLine("13. Deletes the Auto Scaling group.");
        Console.WriteLine("14. Deletes the Amazon EC2 launch template.");
        PressEnter();
    }

    /// <summary>
    /// Display information about the Amazon Ec2 AutoScaling groups passed
    /// in the list of AutoScalingGroup objects.
    /// </summary>
    /// <param name="groups">A list of AutoScalingGroup objects.</param>
    public void DisplayGroupDetails(List<AutoScalingGroup> groups)
    {
        if (groups is null)
            return;

        groups.ForEach(group =>
        {
            Console.WriteLine($"Group name:\t{group.AutoScalingGroupName}");
            Console.WriteLine($"Group created:\t{group.CreatedTime}");
            Console.WriteLine($"Maximum number of instances:\t{group.MaxSize}");
            Console.WriteLine($"Desired number of instances:
\t{group.DesiredCapacity}");
        });
    }

    /// <summary>
    /// Display a message and wait until the user presses enter.
    /// </summary>
    public void PressEnter()
    {
        Console.Write("\nPress <Enter> to continue. ");
        _ = Console.ReadLine();
        Console.WriteLine();
    }
```

```
    /// <summary>
    /// Pad a string with spaces to center it on the console display.
    /// </summary>
    /// <param name="strToCenter">The string to be centered.</param>
    /// <returns>The padded string.</returns>
    public string CenterString(string strToCenter)
    {
        var padAmount = (Console.WindowWidth - strToCenter.Length) / 2;
        var leftPad = new string(' ', padAmount);
        return $"{leftPad}{strToCenter}";
    }

    /// <summary>
    /// Display a line of hyphens, the centered text of the title and another
    /// line of hyphens.
    /// </summary>
    /// <param name="strTitle">The string to be displayed.</param>
    public void DisplayTitle(string strTitle)
    {
        Console.WriteLine(SepBar);
        Console.WriteLine(CenterString(strTitle));
        Console.WriteLine(SepBar);
    }

    /// <summary>
    /// Display a countdown and wait for a number of seconds.
    /// </summary>
    /// <param name="numSeconds">The number of seconds to wait.</param>
    public void WaitABit(int numSeconds, string msg)
    {
        Console.WriteLine(msg);

        // Wait for the requested number of seconds.
        for (int i = numSeconds; i > 0; i--)
        {
            System.Threading.Thread.Sleep(1000);
            Console.Write($"{i}...");
        }

        PressEnter();
    }
}
```

Defina as funções que são chamadas pelo cenário para gerenciar modelos e métricas de lançamento. Essas funções incluem Auto Scaling, Amazon EC2 e ações. CloudWatch

```
namespace AutoScalingActions;

using Amazon.AutoScaling;
using Amazon.AutoScaling.Model;

/// <summary>
/// A class that includes methods to perform Amazon EC2 Auto Scaling
/// actions.
/// </summary>
public class AutoScalingWrapper
{
    private readonly IAmazonAutoScaling _amazonAutoScaling;

    /// <summary>
    /// Constructor for the AutoScalingWrapper class.
    /// </summary>
    /// <param name="amazonAutoScaling">The injected Amazon EC2 Auto Scaling
 client.</param>
    public AutoScalingWrapper(IAmazonAutoScaling amazonAutoScaling)
    {
        _amazonAutoScaling = amazonAutoScaling;
    }


    /// <summary>
    /// Create a new Amazon EC2 Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name to use for the new Auto Scaling
    /// group.</param>
    /// <param name="launchTemplateName">The name of the Amazon EC2 Auto Scaling
    /// launch template to use to create instances in the group.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> CreateAutoScalingGroupAsync(
        string groupName,
        string launchTemplateName,
        string availabilityZone)
    {
```

```
        var templateSpecification = new LaunchTemplateSpecification
        {
            LaunchTemplateName = launchTemplateName,
        };

        var zoneList = new List<string>
            {
                availabilityZone,
            };

        var request = new CreateAutoScalingGroupRequest
        {
            AutoScalingGroupName = groupName,
            AvailabilityZones = zoneList,
            LaunchTemplate = templateSpecification,
            MaxSize = 6,
            MinSize = 1
        };
        try
        {
            var response = await
_amazonAutoScaling.CreateAutoScalingGroupAsync(request);
            Console.WriteLine($"{groupName} Auto Scaling Group created");
            return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
        }
        catch (AlreadyExistsException)
        {
            Console.WriteLine($"{groupName} Auto Scaling Group already exists.");
            return true;
        }
    }



    /// <summary>
    /// Retrieve information about Amazon EC2 Auto Scaling quotas to the
    /// active AWS account.
    /// </summary>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DescribeAccountLimitsAsync()
    {
        var response = await _amazonAutoScaling.DescribeAccountLimitsAsync();
        Console.WriteLine("The maximum number of Auto Scaling groups is " +
response.MaxNumberOfAutoScalingGroups);
```

```
        Console.WriteLine("The current number of Auto Scaling groups is " +
 response.NumberOfAutoScalingGroups);
        return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
    }



    /// <summary>
    /// Retrieve a list of the Amazon EC2 Auto Scaling activities for an
    /// Amazon EC2 Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A list of Amazon EC2 Auto Scaling activities.</returns>
    public async Task<List<Activity>> DescribeScalingActivitiesAsync(
        string groupName)
    {
        var activities = new List<Activity>();
        var scalingActivitiesRequest = new DescribeScalingActivitiesRequest
        {
            AutoScalingGroupName = groupName,
            MaxRecords = 10,
        };

        var response = await
 _amazonAutoScaling.DescribeScalingActivitiesAsync(scalingActivitiesRequest);
        if (response.Activities != null)
        {
            activities = response.Activities;
        }
        return activities;
    }



    /// <summary>
    /// Get data about the instances in an Amazon EC2 Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A list of Amazon EC2 Auto Scaling details.</returns>
    public async Task<List<AutoScalingInstanceDetails>>
 DescribeAutoScalingInstancesAsync(
        string groupName)
```

```
    {
        var groups = await DescribeAutoScalingGroupsAsync(groupName);
        var instanceIds = new List<string>();
        var instanceDetails = new List<AutoScalingInstanceDetails>();
        if (groups != null)
        {
            groups.ForEach(group =>
            {
                if (group.AutoScalingGroupName == groupName && group.Instances !=
 null)
                {
                    group.Instances.ForEach(instance =>
                    {
                        instanceIds.Add(instance.InstanceId);
                    });
                }
            });

            var scalingGroupsRequest = new DescribeAutoScalingInstancesRequest
            {
                MaxRecords = 10,
                InstanceIds = instanceIds,
            };

            var response =
                await _amazonAutoScaling.DescribeAutoScalingInstancesAsync(
                    scalingGroupsRequest);
            if (response.AutoScalingInstances != null)
            {
                instanceDetails = response.AutoScalingInstances;
            }
        }

        return instanceDetails;
    }



    /// <summary>
    /// Retrieve a list of information about Amazon EC2 Auto Scaling groups.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A list of Amazon EC2 Auto Scaling groups.</returns>
```

```
    public async Task<List<AutoScalingGroup>> DescribeAutoScalingGroupsAsync(
        string groupName)
    {
        var groups = new List<AutoScalingGroup>();
        var groupList = new List<string>
            {
                groupName,
            };

        var request = new DescribeAutoScalingGroupsRequest
        {
            AutoScalingGroupNames = groupList,
        };

        var response = await
 _amazonAutoScaling.DescribeAutoScalingGroupsAsync(request);
        if (response.AutoScalingGroups != null)
        {
            groups = response.AutoScalingGroups;
        }

        return groups;
    }


    /// <summary>
    /// Delete an Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteAutoScalingGroupAsync(
        string groupName)
    {
        var deleteAutoScalingGroupRequest = new DeleteAutoScalingGroupRequest
        {
            AutoScalingGroupName = groupName,
            ForceDelete = true,
        };

        var response = await
 _amazonAutoScaling.DeleteAutoScalingGroupAsync(deleteAutoScalingGroupRequest);
        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
```

```
            Console.WriteLine($"You successfully deleted {groupName}");
            return true;
        }

        Console.WriteLine($"Couldn't delete {groupName}.");
        return false;
    }


    /// <summary>
    /// Disable the collection of metric data for an Amazon EC2 Auto Scaling
    /// group.
    /// </summary>
    /// <param name="groupName">The name of the Auto Scaling group.</param>
    /// <returns>A Boolean value that indicates the success or failure of
    /// the operation.</returns>
    public async Task<bool> DisableMetricsCollectionAsync(string groupName)
    {
        var request = new DisableMetricsCollectionRequest
        {
            AutoScalingGroupName = groupName,
        };

        var response = await
_amazonAutoScaling.DisableMetricsCollectionAsync(request);
        return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
    }


    /// <summary>
    /// Enable the collection of metric data for an Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Auto Scaling group.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> EnableMetricsCollectionAsync(string groupName)
    {
        var listMetrics = new List<string>
            {
                "GroupMaxSize",
            };

        var collectionRequest = new EnableMetricsCollectionRequest
        {
            AutoScalingGroupName = groupName,
```

```
            Metrics = listMetrics,
            Granularity = "1Minute",
        };

        var response = await
 _amazonAutoScaling.EnableMetricsCollectionAsync(collectionRequest);
        return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
    }


    /// <summary>
    /// Set the desired capacity of an Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Auto Scaling group.</param>
    /// <param name="desiredCapacity">The desired capacity for the Auto
    /// Scaling group.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> SetDesiredCapacityAsync(
        string groupName,
        int desiredCapacity)
    {
        var capacityRequest = new SetDesiredCapacityRequest
        {
            AutoScalingGroupName = groupName,
            DesiredCapacity = desiredCapacity,
        };

        var response = await
 _amazonAutoScaling.SetDesiredCapacityAsync(capacityRequest);
        Console.WriteLine($"You have set the DesiredCapacity to
 {desiredCapacity}.");

        return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
    }


    /// <summary>
    /// Terminate all instances in the Auto Scaling group in preparation for
    /// deleting the group.
    /// </summary>
    /// <param name="instanceId">The instance Id of the instance to terminate.</
param>
    /// <returns>A Boolean value that indicates the success or failure of
    /// the operation.</returns>
```

```
    public async Task<bool> TerminateInstanceInAutoScalingGroupAsync(
        string instanceId)
    {
        var request = new TerminateInstanceInAutoScalingGroupRequest
        {
            InstanceId = instanceId,
            ShouldDecrementDesiredCapacity = false,
        };

        var response = await
 _amazonAutoScaling.TerminateInstanceInAutoScalingGroupAsync(request);

        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
            Console.WriteLine($"You have terminated the instance: {instanceId}");
            return true;
        }

        Console.WriteLine($"Could not terminate {instanceId}");
        return false;
    }


    /// <summary>
    /// Update the capacity of an Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Auto Scaling group.</param>
    /// <param name="launchTemplateName">The name of the EC2 launch template.</
param>
    /// <param name="maxSize">The maximum number of instances that can be
    /// created for the Auto Scaling group.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> UpdateAutoScalingGroupAsync(
        string groupName,
        string launchTemplateName,
        int maxSize)
    {
        var templateSpecification = new LaunchTemplateSpecification
        {
            LaunchTemplateName = launchTemplateName,
        };

        var groupRequest = new UpdateAutoScalingGroupRequest
        {
```

```
            MaxSize = maxSize,
            AutoScalingGroupName = groupName,
            LaunchTemplate = templateSpecification,
        };

        var response = await
 _amazonAutoScaling.UpdateAutoScalingGroupAsync(groupRequest);
        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
            Console.WriteLine($"You successfully updated the Auto Scaling group
 {groupName}.");
            return true;
        }
        else
        {
            return false;
        }
    }

}


namespace AutoScalingActions;

using Amazon.EC2;
using Amazon.EC2.Model;

public class EC2Wrapper
{
    private readonly IAmazonEC2 _amazonEc2;

    /// <summary>
    /// Constructor for the EC2Wrapper class.
    /// </summary>
    /// <param name="amazonEc2">The injected Amazon EC2 client.</param>
    public EC2Wrapper(IAmazonEC2 amazonEc2)
    {
        _amazonEc2 = amazonEc2;
    }

    /// <summary>
    /// Create a new Amazon EC2 launch template.
    /// </summary>
    /// <param name="imageId">The image Id to use for instances launched
```

```
    /// using the Amazon EC2 launch template.</param>
    /// <param name="instanceType">The type of EC2 instances to create.</param>
    /// <param name="launchTemplateName">The name of the launch template.</param>
    /// <returns>Returns the TemplateID of the new launch template.</returns>
    public async Task<string> CreateLaunchTemplateAsync(
        string imageId,
        string instanceType,
        string launchTemplateName)
    {
        var request = new CreateLaunchTemplateRequest
        {
            LaunchTemplateData = new RequestLaunchTemplateData
            {
                ImageId = imageId,
                InstanceType = instanceType,
            },
            LaunchTemplateName = launchTemplateName,
        };

        var response = await _amazonEc2.CreateLaunchTemplateAsync(request);

        return response.LaunchTemplate.LaunchTemplateId;
    }

    /// <summary>
    /// Delete an Amazon EC2 launch template.
    /// </summary>
    /// <param name="launchTemplateId">The TemplateId of the launch template to
    /// delete.</param>
    /// <returns>The name of the EC2 launch template that was deleted.</returns>
    public async Task<string> DeleteLaunchTemplateAsync(string launchTemplateId)
    {
        var request = new DeleteLaunchTemplateRequest
        {
            LaunchTemplateId = launchTemplateId,
        };

        var response = await _amazonEc2.DeleteLaunchTemplateAsync(request);
        return response.LaunchTemplate.LaunchTemplateName;
    }

    /// <summary>
    /// Retrieve information about an EC2 launch template.
    /// </summary>
```

```
    /// <param name="launchTemplateName">The name of the EC2 launch template.</
param>
    /// <returns>A Boolean value that indicates the success or failure of
    /// the operation.</returns>
    public async Task<bool> DescribeLaunchTemplateAsync(string
 launchTemplateName)
    {
        var request = new DescribeLaunchTemplatesRequest
        {
            LaunchTemplateNames = new List<string> { launchTemplateName, },
        };

        var response = await _amazonEc2.DescribeLaunchTemplatesAsync(request);

        if (response.LaunchTemplates is not null)
        {
            response.LaunchTemplates.ForEach(template =>
            {
                Console.Write($"{template.LaunchTemplateName}\t");
                Console.WriteLine(template.LaunchTemplateId);
            });

            return true;
        }

        return false;
    }

    /// <summary>
    /// Retrieve the availability zones for the current region.
    /// </summary>
    /// <returns>A collection of availability zones.</returns>
    public async Task<List<AvailabilityZone>> ListAvailabilityZonesAsync()
    {
        var response = await _amazonEc2.DescribeAvailabilityZonesAsync(
            new DescribeAvailabilityZonesRequest());

        return response.AvailabilityZones;
    }
}


namespace AutoScalingActions;
```

```
using Amazon.CloudWatch;
using Amazon.CloudWatch.Model;

/// <summary>
/// Contains methods to access Amazon CloudWatch metrics for the
/// Amazon EC2 Auto Scaling basics scenario.
/// </summary>
public class CloudWatchWrapper
{
    private readonly IAmazonCloudWatch _amazonCloudWatch;

    /// <summary>
    /// Constructor for the CloudWatchWrapper.
    /// </summary>
    /// <param name="amazonCloudWatch">The injected CloudWatch client.</param>
    public CloudWatchWrapper(IAmazonCloudWatch amazonCloudWatch)
    {
        _amazonCloudWatch = amazonCloudWatch;
    }

    /// <summary>
    /// Retrieve the metrics information collection for the Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Auto Scaling group.</param>
    /// <returns>A list of Metrics collected for the Auto Scaling group.</
returns>
    public async Task<List<Metric>> GetCloudWatchMetricsAsync(string groupName)
    {
        var metrics = new List<Metric>();
        var filter = new DimensionFilter
        {
            Name = "AutoScalingGroupName",
            Value = $"{groupName}",
        };

        var request = new ListMetricsRequest
        {
            MetricName = "AutoScalingGroupName",
            Dimensions = new List<DimensionFilter> { filter },
            Namespace = "AWS/AutoScaling",
        };

        var response = await _amazonCloudWatch.ListMetricsAsync(request);
        if (response.Metrics != null)
```

```
        {
            metrics = response.Metrics;
        }
        return metrics;
    }

    /// <summary>
    /// Retrieve the metric data collected for an Amazon EC2 Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A list of data points.</returns>
    public async Task<List<Datapoint>> GetMetricStatisticsAsync(string groupName)
    {
        var dataPoints = new List<Datapoint>();
        var metricDimensions = new List<Dimension>
            {
                new Dimension
                {
                    Name = "AutoScalingGroupName",
                    Value = $"{groupName}",
                },
            };

        // The start time will be yesterday.
        var startTime = DateTime.UtcNow.AddDays(-1);

        var request = new GetMetricStatisticsRequest
        {
            MetricName = "AutoScalingGroupName",
            Dimensions = metricDimensions,
            Namespace = "AWS/AutoScaling",
            Period = 60, // 60 seconds.
            Statistics = new List<string>() { "Minimum" },
            StartTimeUtc = startTime,
            EndTimeUtc = DateTime.UtcNow,
        };

        var response = await _amazonCloudWatch.GetMetricStatisticsAsync(request);
        if (response.Datapoints != null)
        {
            dataPoints = response.Datapoints;
        }
```

```
        return dataPoints;
    }

}
```

- Para obter detalhes da API, consulte os tópicos a seguir na Referência da API AWS SDK para .NET .
  - CreateAutoScalingGroup
  - DeleteAutoScalingGroup
  - DescribeAutoScalingGroups
  - DescribeAutoScalingInstances
  - DescribeScalingActivities
  - DisableMetricsCollection
  - EnableMetricsCollection
  - SetDesiredCapacity
  - TerminateInstanceInAutoScalingGroup
  - UpdateAutoScalingGroup

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
//! Routine which demonstrates using an Auto Scaling group
//! to manage Amazon EC2 instances.
/*!
  \sa groupsAndInstancesScenario()
  \param clientConfig: AWS client configuration.
  \return bool: Successful completion.
 */
```

```
bool AwsDoc::AutoScaling::groupsAndInstancesScenario(
        const Aws::Client::ClientConfiguration &clientConfig) {
    Aws::String templateName;
    Aws::EC2::EC2Client ec2Client(clientConfig);

    std::cout << std::setfill('*') << std::setw(ASTERISK_FILL_WIDTH) << " "
              << std::endl;
    std::cout
            << "Welcome to the Amazon Elastic Compute Cloud (Amazon EC2) Auto
 Scaling "
            << "demo for managing groups and instances." << std::endl;
    std::cout << std::setfill('*') << std::setw(ASTERISK_FILL_WIDTH) << " \n"
              << std::endl;

    std::cout << "This example requires an EC2 launch template." << std::endl;
    if (askYesNoQuestion(
            "Would you like to use an existing EC2 launch template (y/n)?  ")) {

        // 1. Specify the name of an existing EC2 launch template.
        templateName = askQuestion(
                "Enter the name of the existing EC2 launch template.  ");

        Aws::EC2::Model::DescribeLaunchTemplatesRequest request;
        request.AddLaunchTemplateNames(templateName);
        Aws::EC2::Model::DescribeLaunchTemplatesOutcome outcome =
                ec2Client.DescribeLaunchTemplates(request);

        if (outcome.IsSuccess()) {
            std::cout << "Validated the EC2 launch template '" << templateName
                      << "' exists by calling DescribeLaunchTemplate." <<
 std::endl;
        }
        else {
            std::cerr << "Error validating the existence of the launch template.
 "
                      << outcome.GetError().GetMessage()
                      << std::endl;
        }
    }
    else { // 2.  Or create a new EC2 launch template.
        templateName = askQuestion("Enter the name for a new EC2 launch template:
 ");

        Aws::EC2::Model::CreateLaunchTemplateRequest request;
```

```
        request.SetLaunchTemplateName(templateName);

        Aws::EC2::Model::RequestLaunchTemplateData requestLaunchTemplateData;

requestLaunchTemplateData.SetInstanceType(EC2_LAUNCH_TEMPLATE_INSTANCE_TYPE);
        requestLaunchTemplateData.SetImageId(EC2_LAUNCH_TEMPLATE_IMAGE_ID);

        request.SetLaunchTemplateData(requestLaunchTemplateData);

        Aws::EC2::Model::CreateLaunchTemplateOutcome outcome =
                ec2Client.CreateLaunchTemplate(request);

        if (outcome.IsSuccess()) {
            std::cout << "The EC2 launch template '" << templateName << " was
created."
                      << std::endl;
        }
        else if (outcome.GetError().GetExceptionName() ==
                 "InvalidLaunchTemplateName.AlreadyExistsException") {
            std::cout << "The EC2 template '" << templateName << "' already
exists"
                      << std::endl;
        }
        else {
            std::cerr << "Error with EC2::CreateLaunchTemplate. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
        }
    }
    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);
    std::cout << "Let's create an Auto Scaling group." << std::endl;
    Aws::String groupName = askQuestion(
            "Enter a name for the Auto Scaling group:  ");
    // 3. Retrieve a list of EC2 Availability Zones.
    Aws::Vector<Aws::EC2::Model::AvailabilityZone> availabilityZones;
    {
        Aws::EC2::Model::DescribeAvailabilityZonesRequest request;

        Aws::EC2::Model::DescribeAvailabilityZonesOutcome outcome =
                ec2Client.DescribeAvailabilityZones(request);

        if (outcome.IsSuccess()) {
            std::cout
```

```
                    << "EC2 instances can be created in the following
Availability Zones:"
                    << std::endl;

            availabilityZones = outcome.GetResult().GetAvailabilityZones();
            for (size_t i = 0; i < availabilityZones.size(); ++i) {
                std::cout << "   " << i + 1 << ".  "
                          << availabilityZones[i].GetZoneName() << std::endl;
            }
        }
        else {
            std::cerr << "Error with EC2::DescribeAvailabilityZones. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
            cleanupResources("", templateName, autoScalingClient, ec2Client);
            return false;
        }
    }

    int availabilityZoneChoice = askQuestionForIntRange(
            "Choose an Availability Zone:  ", 1,
            static_cast<int>(availabilityZones.size()));
    // 4. Create an Auto Scaling group with the specified Availability Zone.
    {
        Aws::AutoScaling::Model::CreateAutoScalingGroupRequest request;
        request.SetAutoScalingGroupName(groupName);
        Aws::Vector<Aws::String> availabilityGroupZones;
        availabilityGroupZones.push_back(
                availabilityZones[availabilityZoneChoice - 1].GetZoneName());
        request.SetAvailabilityZones(availabilityGroupZones);
        request.SetMaxSize(1);
        request.SetMinSize(1);

        Aws::AutoScaling::Model::LaunchTemplateSpecification
launchTemplateSpecification;
        launchTemplateSpecification.SetLaunchTemplateName(templateName);
        request.SetLaunchTemplate(launchTemplateSpecification);

        Aws::AutoScaling::Model::CreateAutoScalingGroupOutcome outcome =
                autoScalingClient.CreateAutoScalingGroup(request);

        if (outcome.IsSuccess()) {
            std::cout << "Created Auto Scaling group '" << groupName << "'..."
                      << std::endl;
```

```
        }
        else if (outcome.GetError().GetErrorType() ==
                 Aws::AutoScaling::AutoScalingErrors::ALREADY_EXISTS_FAULT) {
            std::cout << "Auto Scaling group '" << groupName << "' already
exists."
                      << std::endl;
        }
        else {
            std::cerr << "Error with AutoScaling::CreateAutoScalingGroup. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
            cleanupResources("", templateName, autoScalingClient, ec2Client);
            return false;
        }
    }

    Aws::Vector<Aws::AutoScaling::Model::AutoScalingGroup> autoScalingGroups;
    if (AwsDoc::AutoScaling::describeGroup(groupName, autoScalingGroups,
                                           autoScalingClient)) {
        std::cout << "Here is the Auto Scaling group description." << std::endl;
        if (!autoScalingGroups.empty()) {
            logAutoScalingGroupInfo(autoScalingGroups);
        }
    }
    else {
        cleanupResources(groupName, templateName, autoScalingClient, ec2Client);
        return false;
    }

    std::cout
            << "Waiting for the EC2 instance in the Auto Scaling group to become
active..."
            << std::endl;
    if (!waitForInstances(groupName, autoScalingGroups, autoScalingClient)) {
        cleanupResources(groupName, templateName, autoScalingClient, ec2Client);
        return false;
    }

    bool enableMetrics = askYesNoQuestion(
            "Do you want to collect metrics about the A"
            "Auto Scaling group during this demo (y/n)?  ");
    // 7. Optionally enable metrics collection for the Auto Scaling group.
    if (enableMetrics) {
        Aws::AutoScaling::Model::EnableMetricsCollectionRequest request;
```

```
        request.SetAutoScalingGroupName(groupName);

        request.AddMetrics("GroupMinSize");
        request.AddMetrics("GroupMaxSize");
        request.AddMetrics("GroupDesiredCapacity");
        request.AddMetrics("GroupInServiceInstances");
        request.AddMetrics("GroupTotalInstances");
        request.SetGranularity("1Minute");

        Aws::AutoScaling::Model::EnableMetricsCollectionOutcome outcome =
                autoScalingClient.EnableMetricsCollection(request);
        if (outcome.IsSuccess()) {
            std::cout << "Auto Scaling metrics have been enabled."
                      << std::endl;
        }
        else {
            std::cerr << "Error with AutoScaling::EnableMetricsCollection. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
            cleanupResources(groupName, templateName, autoScalingClient,
ec2Client);
            return false;
        }
    }

    std::cout << "Let's update the maximum number of EC2 instances in '" <<
groupName <<
              "' from 1 to 3." << std::endl;
    askQuestion("Press enter to continue:  ", alwaysTrueTest);
    // 8. Update the Auto Scaling group, setting a new maximum size.
    {
        Aws::AutoScaling::Model::UpdateAutoScalingGroupRequest request;
        request.SetAutoScalingGroupName(groupName);
        request.SetMaxSize(3);

        Aws::AutoScaling::Model::UpdateAutoScalingGroupOutcome outcome =
                autoScalingClient.UpdateAutoScalingGroup(request);

        if (!outcome.IsSuccess()) {
            std::cerr << "Error with AutoScaling::UpdateAutoScalingGroup. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
            cleanupResources(groupName, templateName, autoScalingClient,
ec2Client);
```

```
            return false;
        }
    }

    if (AwsDoc::AutoScaling::describeGroup(groupName, autoScalingGroups,
                                           autoScalingClient)) {
        if (!autoScalingGroups.empty()) {
            const auto &instances = autoScalingGroups[0].GetInstances();
            std::cout
                    << "The group still has one running EC2 instance, but it can
have up to 3.\n"
                    << std::endl;
            logAutoScalingGroupInfo(autoScalingGroups);
        }
        else {
            std::cerr
                    << "No EC2 launch groups were retrieved from DescribeGroup
request."
                    << std::endl;
            cleanupResources(groupName, templateName, autoScalingClient,
ec2Client);
            return false;
        }
    }

    std::cout << "\n" << std::setfill('*') << std::setw(ASTERISK_FILL_WIDTH) <<
"\n"
              << std::endl;
    std::cout << "Let's update the desired capacity in '" << groupName <<
              "' from 1 to 2." << std::endl;
    askQuestion("Press enter to continue:  ", alwaysTrueTest);
    //  9. Update the Auto Scaling group, setting a new desired capacity.
    {
        Aws::AutoScaling::Model::SetDesiredCapacityRequest request;
        request.SetAutoScalingGroupName(groupName);
        request.SetDesiredCapacity(2);

        Aws::AutoScaling::Model::SetDesiredCapacityOutcome outcome =
                autoScalingClient.SetDesiredCapacity(request);

        if (!outcome.IsSuccess()) {
            std::cerr << "Error with AutoScaling::SetDesiredCapacityRequest. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
```

```
            cleanupResources(groupName, templateName, autoScalingClient,
ec2Client);
            return false;
        }
    }

    if (AwsDoc::AutoScaling::describeGroup(groupName, autoScalingGroups,
                                           autoScalingClient)) {
        if (!autoScalingGroups.empty()) {
            std::cout
                    << "Here is the current state of the group." << std::endl;
            logAutoScalingGroupInfo(autoScalingGroups);
        }
        else {
            std::cerr
                    << "No EC2 launch groups were retrieved from DescribeGroup
request."
                    << std::endl;
            cleanupResources(groupName, templateName, autoScalingClient,
ec2Client);
            return false;
        }
    }

    std::cout << "Waiting for the new EC2 instance to start..." << std::endl;
    waitForInstances(groupName, autoScalingGroups, autoScalingClient);

    std::cout << "\n" << std::setfill('*') << std::setw(ASTERISK_FILL_WIDTH) <<
"\n"
              << std::endl;

    std::cout << "Let's terminate one of the EC2 instances in " << groupName <<
"."
              << std::endl;
    std::cout << "Because the desired capacity is 2, another EC2 instance will
start "
              << "to replace the terminated EC2 instance."
              << std::endl;
    std::cout << "The currently running EC2 instances are:" << std::endl;

    if (autoScalingGroups.empty()) {
        std::cerr << "Error describing groups. No groups returned." << std::endl;
        cleanupResources(groupName, templateName, autoScalingClient, ec2Client);
        return false;
```

```
    }

    int instanceNumber = 1;
    Aws::Vector<Aws::String> instanceIDs = instancesToInstanceIDs(
            autoScalingGroups[0].GetInstances());
    for (const Aws::String &instanceID: instanceIDs) {
        std::cout << "   " << instanceNumber << ". " << instanceID << std::endl;
        ++instanceNumber;
    }

    instanceNumber = askQuestionForIntRange("Which EC2 instance do you want to
stop? ",
                                            1,

static_cast<int>(instanceIDs.size()));

    // 10. Terminate an EC2 instance in the Auto Scaling group.
    {
        Aws::AutoScaling::Model::TerminateInstanceInAutoScalingGroupRequest
request;
        request.SetInstanceId(instanceIDs[instanceNumber - 1]);
        request.SetShouldDecrementDesiredCapacity(false);

        Aws::AutoScaling::Model::TerminateInstanceInAutoScalingGroupOutcome
outcome =
                autoScalingClient.TerminateInstanceInAutoScalingGroup(request);

        if (outcome.IsSuccess()) {
            std::cout << "Waiting for EC2 instance with ID '"
                      << instanceIDs[instanceNumber - 1] << "' to terminate..."
                      << std::endl;
        }
        else {
            std::cerr << "Error with
AutoScaling::TerminateInstanceInAutoScalingGroup. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
            cleanupResources(groupName, templateName, autoScalingClient,
ec2Client);
            return false;
        }
    }

    waitForInstances(groupName, autoScalingGroups, autoScalingClient);
```

```
   std::cout << "\n" << std::setfill('*') << std::setw(ASTERISK_FILL_WIDTH) <<
"\n"
             << std::endl;
   std::cout << "Let's get a report of scaling activities for EC2 launch group
'"
             << groupName << "'."
             << std::endl;
   askQuestion("Press enter to continue:  ", alwaysTrueTest);
   // 11. Get a description of activities for the Auto Scaling group.
   {
       Aws::AutoScaling::Model::DescribeScalingActivitiesRequest request;
       request.SetAutoScalingGroupName(groupName);

       Aws::Vector<Aws::AutoScaling::Model::Activity> allActivities;
       Aws::String nextToken; // Used for pagination;
       do {
           if (!nextToken.empty()) {
               request.SetNextToken(nextToken);
           }
           Aws::AutoScaling::Model::DescribeScalingActivitiesOutcome outcome =
                   autoScalingClient.DescribeScalingActivities(request);

           if (outcome.IsSuccess()) {
               const Aws::Vector<Aws::AutoScaling::Model::Activity> &activities
=
                       outcome.GetResult().GetActivities();
               allActivities.insert(allActivities.end(), activities.begin(),
activities.end());
               nextToken  = outcome.GetResult().GetNextToken();
           }
           else {
               std::cerr << "Error with AutoScaling::DescribeScalingActivities.
"
                         << outcome.GetError().GetMessage()
                         << std::endl;
               cleanupResources(groupName, templateName, autoScalingClient,
ec2Client);
               return false;
           }
       } while (!nextToken.empty());

       std::cout << "Found " << allActivities.size() << " activities."
                 << std::endl;
```

```
        std::cout << "Activities are ordered with the most recent first."
                  << std::endl;
        for (const Aws::AutoScaling::Model::Activity &activity: allActivities) {
            std::cout << activity.GetDescription() << std::endl;
            std::cout << activity.GetDetails() << std::endl;
        }
    }

    if (enableMetrics) {
        if (!logAutoScalingMetrics(groupName, clientConfig)) {
            cleanupResources(groupName, templateName, autoScalingClient,
 ec2Client);
            return false;
        }
    }

    std::cout << "Let's  clean up." << std::endl;
    askQuestion("Press enter to continue:  ", alwaysTrueTest);

    // 13. Disable metrics collection if enabled.
    if (enableMetrics) {
        Aws::AutoScaling::Model::DisableMetricsCollectionRequest request;
        request.SetAutoScalingGroupName(groupName);

        Aws::AutoScaling::Model::DisableMetricsCollectionOutcome outcome =
                autoScalingClient.DisableMetricsCollection(request);

        if (outcome.IsSuccess()) {
            std::cout << "Metrics collection has been disabled." << std::endl;
        }
        else {
            std::cerr << "Error with AutoScaling::DisableMetricsCollection. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
            cleanupResources(groupName, templateName, autoScalingClient,
 ec2Client);
            return false;
        }
    }

    return cleanupResources(groupName, templateName, autoScalingClient,
 ec2Client);
}
```

```
//! Routine which waits for EC2 instances in an Auto Scaling group to
//! complete startup or shutdown.
/*!
 \sa waitForInstances()
 \param groupName: An Auto Scaling group name.
 \param autoScalingGroups: Vector to receive 'AutoScalingGroup' records.
 \param client: 'AutoScalingClient' instance.
 \return bool: Successful completion.
 */
bool AwsDoc::AutoScaling::waitForInstances(const Aws::String &groupName,

 Aws::Vector<Aws::AutoScaling::Model::AutoScalingGroup> &autoScalingGroups,
                                            const
 Aws::AutoScaling::AutoScalingClient &client) {
    bool ready = false;
    const std::vector<Aws::String> READY_STATES = {"InService", "Terminated"};

    int count = 0;
    int desiredCapacity = 0;
    std::this_thread::sleep_for(std::chrono::seconds(4));
    while (!ready) {
        if (WAIT_FOR_INSTANCES_TIMEOUT < count) {
            std::cerr << "Wait for instance timed out." << std::endl;
            return false;
        }

        std::this_thread::sleep_for(std::chrono::seconds(1));
        ++count;
        if (!describeGroup(groupName, autoScalingGroups, client)) {
            return false;
        }
        Aws::Vector<Aws::String> instanceIDs;
        if (!autoScalingGroups.empty()) {
            instanceIDs =
 instancesToInstanceIDs(autoScalingGroups[0].GetInstances());
            desiredCapacity = autoScalingGroups[0].GetDesiredCapacity();
        }

        if (instanceIDs.empty()) {
            if (desiredCapacity == 0) {
                break;
            }
            else {
                if ((count % 5) == 0) {
```

```
                    std::cout << "No instance IDs returned for group." <<
std::endl;
                }

                continue;
            }
        }

        // 6.  Check lifecycle state of the instances using
DescribeAutoScalingInstances.
        Aws::AutoScaling::Model::DescribeAutoScalingInstancesRequest request;
        request.SetInstanceIds(instanceIDs);

        Aws::AutoScaling::Model::DescribeAutoScalingInstancesOutcome outcome =
                client.DescribeAutoScalingInstances(request);

        if (outcome.IsSuccess()) {
            const
Aws::Vector<Aws::AutoScaling::Model::AutoScalingInstanceDetails>
&instancesDetails =
                    outcome.GetResult().GetAutoScalingInstances();
            ready = instancesDetails.size() >= desiredCapacity;
            for (const Aws::AutoScaling::Model::AutoScalingInstanceDetails
&details: instancesDetails) {
                if (!stringInVector(details.GetLifecycleState(), READY_STATES)) {
                    ready = false;
                    break;
                }
            }
            // Log the status while waiting.
            if (((count % 5) == 1) || ready) {
                logInstancesLifecycleState(instancesDetails);
            }
        }
        else {
            std::cerr << "Error with AutoScaling::DescribeAutoScalingInstances. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
            return false;
        }
    }

    if (!describeGroup(groupName, autoScalingGroups, client)) {
        return false;
```

```
    }

    return true;
}

//! Routine to cleanup resources created in 'groupsAndInstancesScenario'.
/*!
 \sa cleanupResources()
 \param groupName: Optional Auto Scaling group name.
 \param templateName: Optional EC2 launch template name.
 \param autoScalingClient: 'AutoScalingClient' instance.
 \param ec2Client: 'EC2Client' instance.
\return bool: Successful completion.
 */
bool AwsDoc::AutoScaling::cleanupResources(const Aws::String &groupName,
                                           const Aws::String &templateName,
                                           const
 Aws::AutoScaling::AutoScalingClient &autoScalingClient,
                                           const Aws::EC2::EC2Client &ec2Client)
 {
    bool result = true;

    // 14. Delete the Auto Scaling group.
    if (!groupName.empty() &&
        (askYesNoQuestion(
                Aws::String("Delete the Auto Scaling group '") + groupName +
                "'  (y/n)?"))) {
        {
            Aws::AutoScaling::Model::UpdateAutoScalingGroupRequest request;
            request.SetAutoScalingGroupName(groupName);
            request.SetMinSize(0);
            request.SetDesiredCapacity(0);

            Aws::AutoScaling::Model::UpdateAutoScalingGroupOutcome outcome =
                    autoScalingClient.UpdateAutoScalingGroup(request);

            if (outcome.IsSuccess()) {
                std::cout
                        << "The minimum size and desired capacity of the Auto
 Scaling group "
                        << "was set to zero before terminating the instances."
                        << std::endl;
            }
            else {
```

```
                std::cerr << "Error with AutoScaling::UpdateAutoScalingGroup. "
                          << outcome.GetError().GetMessage() << std::endl;
                result = false;
            }
        }

        Aws::Vector<Aws::AutoScaling::Model::AutoScalingGroup> autoScalingGroups;
        if (AwsDoc::AutoScaling::describeGroup(groupName, autoScalingGroups,
                                               autoScalingClient)) {
            if (!autoScalingGroups.empty()) {
                Aws::Vector<Aws::String> instanceIDs = instancesToInstanceIDs(
                        autoScalingGroups[0].GetInstances());
                for (const Aws::String &instanceID: instanceIDs) {

Aws::AutoScaling::Model::TerminateInstanceInAutoScalingGroupRequest request;
                    request.SetInstanceId(instanceID);
                    request.SetShouldDecrementDesiredCapacity(true);


Aws::AutoScaling::Model::TerminateInstanceInAutoScalingGroupOutcome outcome =

autoScalingClient.TerminateInstanceInAutoScalingGroup(
                                    request);

                    if (outcome.IsSuccess()) {
                        std::cout << "Initiating termination of EC2 instance '"
                                  << instanceID << "'." << std::endl;
                    }
                    else {
                        std::cerr
                                << "Error with
AutoScaling::TerminateInstanceInAutoScalingGroup. "
                                << outcome.GetError().GetMessage() << std::endl;
                        result = false;
                    }
                }
            }

            std::cout
                    << "Waiting for the EC2 instances to terminate before
deleting the "
                    << "Auto Scaling group..." << std::endl;
            waitForInstances(groupName, autoScalingGroups, autoScalingClient);
        }
```

```
        {
            Aws::AutoScaling::Model::DeleteAutoScalingGroupRequest request;
            request.SetAutoScalingGroupName(groupName);

            Aws::AutoScaling::Model::DeleteAutoScalingGroupOutcome outcome =
                    autoScalingClient.DeleteAutoScalingGroup(request);

            if (outcome.IsSuccess()) {
                std::cout << "Auto Scaling group '" << groupName << "' was
deleted."
                          << std::endl;
            }
            else {
                std::cerr << "Error with AutoScaling::DeleteAutoScalingGroup. "
                          << outcome.GetError().GetMessage()
                          << std::endl;
                result = false;
            }
        }
    }

    // 15. Delete the EC2 launch template.
    if (!templateName.empty() && (askYesNoQuestion(
            Aws::String("Delete the EC2 launch template '") + templateName +
            "' (y/n)?"))) {
        Aws::EC2::Model::DeleteLaunchTemplateRequest request;
        request.SetLaunchTemplateName(templateName);

        Aws::EC2::Model::DeleteLaunchTemplateOutcome outcome =
                ec2Client.DeleteLaunchTemplate(request);

        if (outcome.IsSuccess()) {
            std::cout << "EC2 launch template '" << templateName << "' was
deleted."
                      << std::endl;
        }
        else {
            std::cerr << "Error with EC2::DeleteLaunchTemplate. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
            result = false;
        }
    }
```

```
    return result;
}

//! Routine which retrieves Auto Scaling group descriptions.
/*!
 \sa describeGroup()
 \param groupName: An Auto Scaling group name.
 \param autoScalingGroups: Vector to receive 'AutoScalingGroup' records.
 \param client: 'AutoScalingClient' instance.
 \return bool: Successful completion.
 */
bool AwsDoc::AutoScaling::describeGroup(const Aws::String &groupName,

 Aws::Vector<Aws::AutoScaling::Model::AutoScalingGroup> &autoScalingGroup,
                                         const Aws::AutoScaling::AutoScalingClient
 &client) {
    // 5. Retrieve a description of the Auto Scaling group.
    Aws::AutoScaling::Model::DescribeAutoScalingGroupsRequest request;
    Aws::Vector<Aws::String> groupNames;
    groupNames.push_back(groupName);
    request.SetAutoScalingGroupNames(groupNames);

    Aws::AutoScaling::Model::DescribeAutoScalingGroupsOutcome outcome =
            client.DescribeAutoScalingGroups(request);

    if (outcome.IsSuccess()) {
        autoScalingGroup = outcome.GetResult().GetAutoScalingGroups();
    }
    else {
        std::cerr << "Error with AutoScaling::DescribeAutoScalingGroups. "
                  << outcome.GetError().GetMessage()
                  << std::endl;
    }

    return outcome.IsSuccess();
}
```

- Para obter detalhes da API, consulte os tópicos a seguir na Referência da API AWS SDK para C++ .
  - CreateAutoScalingGroup

- DeleteAutoScalingGroup
- DescribeAutoScalingGroups
- DescribeAutoScalingInstances
- DescribeScalingActivities
- DisableMetricsCollection
- EnableMetricsCollection
- SetDesiredCapacity
- TerminateInstanceInAutoScalingGroup
- UpdateAutoScalingGroup

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
/**
 * Before running this SDK for Java (v2) code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 *
 * In addition, create a launch template. For more information, see the
 * following topic:
 *
 * https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-launch-
templates.html#create-launch-template
 *
 * This code example performs the following operations:
 * 1. Creates an Auto Scaling group using an AutoScalingWaiter.
 * 2. Gets a specific Auto Scaling group and returns an instance Id value.
```

```
 * 3. Describes Auto Scaling with the Id value.
 * 4. Enables metrics collection.
 * 5. Update an Auto Scaling group.
 * 6. Describes Account details.
 * 7. Describe account details"
 * 8. Updates an Auto Scaling group to use an additional instance.
 * 9. Gets the specific Auto Scaling group and gets the number of instances.
 * 10. List the scaling activities that have occurred for the group.
 * 11. Terminates an instance in the Auto Scaling group.
 * 12. Stops the metrics collection.
 * 13. Deletes the Auto Scaling group.
 */

public class AutoScalingScenario {
    public static final String DASHES = new String(new char[80]).replace("\0",
 "-");
    private static final String ROLES_STACK = "MyCdkAutoScaleStack";
    public static void main(String[] args) throws InterruptedException {
        final String usage = """

                Usage:
                    <groupName>

                Where:
                    groupName - The name of the Auto Scaling group.
                """;

        String groupName = "MyAutoScalingGroup2";
        AutoScalingClient autoScalingClient = AutoScalingClient.builder()
                .region(Region.US_WEST_2)
                .build();

        Ec2Client ec2 = Ec2Client.builder()
                .region(Region.US_WEST_2)
                .build();

        System.out.println(DASHES);
        System.out.println("Welcome to the Amazon EC2 Auto Scaling example
 scenario.");
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("First, we will create a launch template using a
 CloudFormation script");
```

```
        CloudFormationHelper.deployCloudFormationStack(ROLES_STACK);
        Map<String, String> stackOutputs =
CloudFormationHelper.getStackOutputsAsync(ROLES_STACK).join();
        String launchTemplateName = stackOutputs.get("LaunchTemplateNameOutput");
        String vpcZoneId = getVPC(ec2);
        updateTemlate(ec2, launchTemplateName );
        System.out.println("The VPC zone id created by the CloudFormation stack
is"+vpcZoneId);

        System.out.println("1. Create an Auto Scaling group named " + groupName);
        createAutoScalingGroup(autoScalingClient, ec2, groupName,
launchTemplateName, vpcZoneId);

        System.out.println(
                "Wait 1 min for the resources, including the instance. Otherwise,
an empty instance Id is returned");
        Thread.sleep(60000);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("2. Get Auto Scale group Id value");
        String instanceId = getSpecificAutoScalingGroups(autoScalingClient,
groupName);
        if (instanceId.compareTo("") == 0) {
            System.out.println("Error - no instance Id value");
            System.exit(1);
        } else {
            System.out.println("The instance Id value is " + instanceId);
        }
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("3. Describe Auto Scaling with the Id value " +
instanceId);
        describeAutoScalingInstance(autoScalingClient, instanceId);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("4. Enable metrics collection " + instanceId);
        enableMetricsCollection(autoScalingClient, groupName);
        System.out.println(DASHES);

        System.out.println(DASHES);
```

```
        System.out.println("5. Update an Auto Scaling group to update max size to
3");
        updateAutoScalingGroup(autoScalingClient, groupName, launchTemplateName);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("6. Describe Auto Scaling groups");
        describeAutoScalingGroups(autoScalingClient, groupName);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("7. Describe account details");
        describeAccountLimits(autoScalingClient);
        System.out.println(
                "Wait 1 min for the resources, including the instance. Otherwise,
an empty instance Id is returned");
        Thread.sleep(60000);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("8. Set desired capacity to 2");
        setDesiredCapacity(autoScalingClient, groupName);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("9. Get the two instance Id values and state");
        getSpecificAutoScalingGroups(autoScalingClient, groupName);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("10. List the scaling activities that have occurred
for the group");
        describeScalingActivities(autoScalingClient, groupName);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("11. Terminate an instance in the Auto Scaling
group");
        terminateInstanceInAutoScalingGroup(autoScalingClient, instanceId);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("12. Stop the metrics collection");
        disableMetricsCollection(autoScalingClient, groupName);
```

```
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("13. Delete the Auto Scaling group and cloud formation
resources");
        CloudFormationHelper.destroyCloudFormationStack(ROLES_STACK);
        deleteAutoScalingGroup(autoScalingClient, groupName);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("The Scenario has successfully completed.");
        System.out.println(DASHES);

        autoScalingClient.close();
    }

    public static String getVPC(Ec2Client ec2) {
        try {
            DescribeVpcsRequest request = DescribeVpcsRequest.builder()
                    .filters(f -> f.name("isDefault").values("true"))
                    .build();

            DescribeVpcsResponse response = ec2.describeVpcs(request);

            if (!response.vpcs().isEmpty()) {
                Vpc defaultVpc = response.vpcs().get(0);
                System.out.println("Default VPC ID: " + defaultVpc.vpcId());
                return defaultVpc.vpcId();
            } else {
                System.out.println("No default VPC found.");
                return null; // Return null if no default VPC is found
            }

        } catch (Ec2Exception e) {
            System.err.println("EC2 error: " +
e.awsErrorDetails().errorMessage());
            return null; // Return null in case of an error
        }
    }


    public static void updateTemlate(Ec2Client ec2, String launchTemplateName ) {
        // Step 1: Create new launch template version
        String newAmiId = "ami-0025f0db847eb6254";
```

```
        RequestLaunchTemplateData launchTemplateData =
RequestLaunchTemplateData.builder()
                .imageId(newAmiId)
                .build();

        CreateLaunchTemplateVersionRequest createVersionRequest =
CreateLaunchTemplateVersionRequest.builder()
                .launchTemplateName(launchTemplateName)
                .versionDescription("Updated with valid AMI")
                .sourceVersion("1")
                .launchTemplateData(launchTemplateData)
                .build();

        CreateLaunchTemplateVersionResponse createResponse =
ec2.createLaunchTemplateVersion(createVersionRequest);
        int newVersionNumber =
createResponse.launchTemplateVersion().versionNumber().intValue();

        // Step 2: Modify default version
        ModifyLaunchTemplateRequest modifyRequest =
ModifyLaunchTemplateRequest.builder()
                .launchTemplateName(launchTemplateName)
                .defaultVersion(String.valueOf(newVersionNumber))
                .build();

        ec2.modifyLaunchTemplate(modifyRequest);
        System.out.println("Updated launch template to version " +
newVersionNumber + " with AMI " + newAmiId);
    }


    public static void describeScalingActivities(AutoScalingClient
autoScalingClient, String groupName) {
        try {
            DescribeScalingActivitiesRequest scalingActivitiesRequest =
DescribeScalingActivitiesRequest.builder()
                    .autoScalingGroupName(groupName)
                    .maxRecords(10)
                    .build();

            DescribeScalingActivitiesResponse response = autoScalingClient
                    .describeScalingActivities(scalingActivitiesRequest);
            List<Activity> activities = response.activities();
            for (Activity activity : activities) {
```

```
                System.out.println("The activity Id is " +
 activity.activityId());
                System.out.println("The activity details are " +
 activity.details());
            }

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }

    public static void setDesiredCapacity(AutoScalingClient autoScalingClient,
 String groupName) {
        try {
            SetDesiredCapacityRequest capacityRequest =
 SetDesiredCapacityRequest.builder()
                    .autoScalingGroupName(groupName)
                    .desiredCapacity(2)
                    .build();

            autoScalingClient.setDesiredCapacity(capacityRequest);
            System.out.println("You have set the DesiredCapacity to 2");

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }

    public static void createAutoScalingGroup(AutoScalingClient
 autoScalingClient,
                                               Ec2Client ec2Client,
                                               String groupName,
                                               String launchTemplateName,
                                               String vpcId) {
        try {
            // Step 1: Get one subnet ID in the given VPC
            DescribeSubnetsRequest subnetRequest =
 DescribeSubnetsRequest.builder()
                    .filters(Filter.builder().name("vpc-
id").values(vpcId).build())
                    .build();
```

```
            DescribeSubnetsResponse subnetResponse =
ec2Client.describeSubnets(subnetRequest);

            if (subnetResponse.subnets().isEmpty()) {
                throw new RuntimeException("No subnets found in VPC: " + vpcId);
            }

            String subnetId = subnetResponse.subnets().get(0).subnetId(); // Use
first subnet
            System.out.println("Using subnet: " + subnetId);

            // Step 2: Create launch template reference
            LaunchTemplateSpecification templateSpecification =
LaunchTemplateSpecification.builder()
                    .launchTemplateName(launchTemplateName)
                    .build();

            // Step 3: Create Auto Scaling group
            CreateAutoScalingGroupRequest request =
CreateAutoScalingGroupRequest.builder()
                    .autoScalingGroupName(groupName)
                    .launchTemplate(templateSpecification)
                    .minSize(1)
                    .maxSize(1)
                    .vpcZoneIdentifier(subnetId)  // Correct: subnet ID, not VPC
ID
                    .build();

            autoScalingClient.createAutoScalingGroup(request);

            // Step 4: Wait until group is created
            AutoScalingWaiter waiter = autoScalingClient.waiter();
            DescribeAutoScalingGroupsRequest groupsRequest =
DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();

            WaiterResponse<DescribeAutoScalingGroupsResponse> waiterResponse =
                    waiter.waitUntilGroupExists(groupsRequest);

            waiterResponse.matched().response().ifPresent(System.out::println);
            System.out.println("Auto Scaling Group created");

        } catch (Ec2Exception | AutoScalingException e) {
```

```
            System.err.println("Error: " + e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }

    public static void describeAutoScalingInstance(AutoScalingClient
autoScalingClient, String id) {
        try {
            DescribeAutoScalingInstancesRequest
describeAutoScalingInstancesRequest = DescribeAutoScalingInstancesRequest
                    .builder()
                    .instanceIds(id)
                    .build();

            DescribeAutoScalingInstancesResponse response = autoScalingClient

.describeAutoScalingInstances(describeAutoScalingInstancesRequest);
            List<AutoScalingInstanceDetails> instances =
response.autoScalingInstances();
            for (AutoScalingInstanceDetails instance : instances) {
                System.out.println("The instance lifecycle state is: " +
instance.lifecycleState());
            }

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }

    public static void describeAutoScalingGroups(AutoScalingClient
autoScalingClient, String groupName) {
        try {
            DescribeAutoScalingGroupsRequest groupsRequest =
DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .maxRecords(10)
                    .build();

            DescribeAutoScalingGroupsResponse response =
autoScalingClient.describeAutoScalingGroups(groupsRequest);
            List<AutoScalingGroup> groups = response.autoScalingGroups();
            for (AutoScalingGroup group : groups) {
```

```
                System.out.println("*** The service to use for the health checks:
" + group.healthCheckType());
            }

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }

    public static String getSpecificAutoScalingGroups(AutoScalingClient
autoScalingClient, String groupName) {
        try {
            String instanceId = "";
            DescribeAutoScalingGroupsRequest scalingGroupsRequest =
DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();

            DescribeAutoScalingGroupsResponse response = autoScalingClient
                    .describeAutoScalingGroups(scalingGroupsRequest);
            List<AutoScalingGroup> groups = response.autoScalingGroups();
            for (AutoScalingGroup group : groups) {
                System.out.println("The group name is " +
group.autoScalingGroupName());
                System.out.println("The group ARN is " +
group.autoScalingGroupARN());
                List<Instance> instances = group.instances();

                for (Instance instance : instances) {
                    instanceId = instance.instanceId();
                    System.out.println("The instance id is " + instanceId);
                    System.out.println("The lifecycle state is " +
instance.lifecycleState());
                }
            }

            return instanceId;
        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
        return "";
    }
```

```
   public static void enableMetricsCollection(AutoScalingClient
autoScalingClient, String groupName) {
       try {
           EnableMetricsCollectionRequest collectionRequest =
EnableMetricsCollectionRequest.builder()
                   .autoScalingGroupName(groupName)
                   .metrics("GroupMaxSize")
                   .granularity("1Minute")
                   .build();

           autoScalingClient.enableMetricsCollection(collectionRequest);
           System.out.println("The enable metrics collection operation was
successful");

       } catch (AutoScalingException e) {
           System.err.println(e.awsErrorDetails().errorMessage());
           System.exit(1);
       }
   }

   public static void disableMetricsCollection(AutoScalingClient
autoScalingClient, String groupName) {
       try {
           DisableMetricsCollectionRequest disableMetricsCollectionRequest =
DisableMetricsCollectionRequest.builder()
                   .autoScalingGroupName(groupName)
                   .metrics("GroupMaxSize")
                   .build();


autoScalingClient.disableMetricsCollection(disableMetricsCollectionRequest);
           System.out.println("The disable metrics collection operation was
successful");

       } catch (AutoScalingException e) {
           System.err.println(e.awsErrorDetails().errorMessage());
           System.exit(1);
       }
   }

   public static void describeAccountLimits(AutoScalingClient autoScalingClient)
{
       try {
```

```
            DescribeAccountLimitsResponse response =
autoScalingClient.describeAccountLimits();
            System.out.println("The max number of auto scaling groups is " +
response.maxNumberOfAutoScalingGroups());
            System.out.println("The current number of auto scaling groups is " +
response.numberOfAutoScalingGroups());

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }

    public static void updateAutoScalingGroup(AutoScalingClient
autoScalingClient, String groupName,
            String launchTemplateName) {
        try {
            AutoScalingWaiter waiter = autoScalingClient.waiter();
            LaunchTemplateSpecification templateSpecification =
LaunchTemplateSpecification.builder()
                    .launchTemplateName(launchTemplateName)
                    .build();

            UpdateAutoScalingGroupRequest groupRequest =
UpdateAutoScalingGroupRequest.builder()
                    .maxSize(3)
                    .autoScalingGroupName(groupName)
                    .launchTemplate(templateSpecification)
                    .build();

            autoScalingClient.updateAutoScalingGroup(groupRequest);
            DescribeAutoScalingGroupsRequest groupsRequest =
DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();

            WaiterResponse<DescribeAutoScalingGroupsResponse> waiterResponse =
waiter
                    .waitUntilGroupInService(groupsRequest);
            waiterResponse.matched().response().ifPresent(System.out::println);
            System.out.println("You successfully updated the auto scaling group
" + groupName);

        } catch (AutoScalingException e) {
```

```
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }

    public static void terminateInstanceInAutoScalingGroup(AutoScalingClient
 autoScalingClient, String instanceId) {
        try {
            TerminateInstanceInAutoScalingGroupRequest request =
 TerminateInstanceInAutoScalingGroupRequest.builder()
                    .instanceId(instanceId)
                    .shouldDecrementDesiredCapacity(false)
                    .build();

            autoScalingClient.terminateInstanceInAutoScalingGroup(request);
            System.out.println("You have terminated instance " + instanceId);

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }

    public static void deleteAutoScalingGroup(AutoScalingClient
 autoScalingClient, String groupName) {
        try {
            DeleteAutoScalingGroupRequest deleteAutoScalingGroupRequest =
 DeleteAutoScalingGroupRequest.builder()
                    .autoScalingGroupName(groupName)
                    .forceDelete(true)
                    .build();


 autoScalingClient.deleteAutoScalingGroup(deleteAutoScalingGroupRequest);
            System.out.println("You successfully deleted " + groupName);

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- Para obter detalhes da API, consulte os tópicos a seguir na Referência da API AWS SDK for Java 2.x .
  - CreateAutoScalingGroup
  - DeleteAutoScalingGroup
  - DescribeAutoScalingGroups
  - DescribeAutoScalingInstances
  - DescribeScalingActivities
  - DisableMetricsCollection
  - EnableMetricsCollection
  - SetDesiredCapacity
  - TerminateInstanceInAutoScalingGroup
  - UpdateAutoScalingGroup

Kotlin

SDK para Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun main(args: Array<String>) {
    val usage = """
    Usage:
        <groupName> <launchTemplateName> <serviceLinkedRoleARN> <vpcZoneId>

    Where:
        groupName - The name of the Auto Scaling group.
        launchTemplateName - The name of the launch template.
        serviceLinkedRoleARN - The Amazon Resource Name (ARN) of the service-
linked role that the Auto Scaling group uses.
        vpcZoneId - A subnet Id for a virtual private cloud (VPC) where instances
 in the Auto Scaling group can be created.
    """
```

```
    if (args.size != 4) {
        println(usage)
        exitProcess(1)
    }

    val groupName = args[0]
    val launchTemplateName = args[1]
    val serviceLinkedRoleARN = args[2]
    val vpcZoneId = args[3]

    println("**** Create an Auto Scaling group named $groupName")
    createAutoScalingGroup(groupName, launchTemplateName, serviceLinkedRoleARN,
vpcZoneId)

    println("Wait 1 min for the resources, including the instance. Otherwise, an
empty instance Id is returned")
    delay(60000)

    val instanceId = getSpecificAutoScaling(groupName)
    if (instanceId.compareTo("") == 0) {
        println("Error - no instance Id value")
        exitProcess(1)
    } else {
        println("The instance Id value is $instanceId")
    }

    println("**** Describe Auto Scaling with the Id value $instanceId")
    describeAutoScalingInstance(instanceId)

    println("**** Enable metrics collection $instanceId")
    enableMetricsCollection(groupName)

    println("**** Update an Auto Scaling group to maximum size of 3")
    updateAutoScalingGroup(groupName, launchTemplateName, serviceLinkedRoleARN)

    println("**** Describe all Auto Scaling groups to show the current state of
the groups")
    describeAutoScalingGroups(groupName)

    println("**** Describe account details")
    describeAccountLimits()

    println("Wait 1 min for the resources, including the instance. Otherwise, an
empty instance Id is returned")
```

```
    delay(60000)

    println("**** Set desired capacity to 2")
    setDesiredCapacity(groupName)

    println("**** Get the two instance Id values and state")
    getAutoScalingGroups(groupName)

    println("**** List the scaling activities that have occurred for the group")
    describeScalingActivities(groupName)

    println("**** Terminate an instance in the Auto Scaling group")
    terminateInstanceInAutoScalingGroup(instanceId)

    println("**** Stop the metrics collection")
    disableMetricsCollection(groupName)

    println("**** Delete the Auto Scaling group")
    deleteSpecificAutoScalingGroup(groupName)
}

suspend fun describeAutoScalingGroups(groupName: String) {
    val groupsReques =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
            maxRecords = 10
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        val response = autoScalingClient.describeAutoScalingGroups(groupsReques)
        response.autoScalingGroups?.forEach { group ->
            println("The service to use for the health checks:
 ${group.healthCheckType}")
        }
    }
}

suspend fun disableMetricsCollection(groupName: String) {
    val disableMetricsCollectionRequest =
        DisableMetricsCollectionRequest {
            autoScalingGroupName = groupName
            metrics = listOf("GroupMaxSize")
        }
```

```
    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->

 autoScalingClient.disableMetricsCollection(disableMetricsCollectionRequest)
        println("The disable metrics collection operation was successful")
    }
}

suspend fun describeScalingActivities(groupName: String?) {
    val scalingActivitiesRequest =
        DescribeScalingActivitiesRequest {
            autoScalingGroupName = groupName
            maxRecords = 10
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        val response =
 autoScalingClient.describeScalingActivities(scalingActivitiesRequest)
        response.activities?.forEach { activity ->
            println("The activity Id is ${activity.activityId}")
            println("The activity details are ${activity.details}")
        }
    }
}

suspend fun getAutoScalingGroups(groupName: String) {
    val scalingGroupsRequest =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        val response =
 autoScalingClient.describeAutoScalingGroups(scalingGroupsRequest)
        response.autoScalingGroups?.forEach { group ->
            println("The group name is ${group.autoScalingGroupName}")
            println("The group ARN is ${group.autoScalingGroupArn}")
            group.instances?.forEach { instance ->
                println("The instance id is ${instance.instanceId}")
                println("The lifecycle state is " + instance.lifecycleState)
            }
        }
    }
}
```

```
suspend fun setDesiredCapacity(groupName: String) {
    val capacityRequest =
        SetDesiredCapacityRequest {
            autoScalingGroupName = groupName
            desiredCapacity = 2
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.setDesiredCapacity(capacityRequest)
        println("You set the DesiredCapacity to 2")
    }
}

suspend fun updateAutoScalingGroup(
    groupName: String,
    launchTemplateNameVal: String,
    serviceLinkedRoleARNVal: String,
) {
    val templateSpecification =
        LaunchTemplateSpecification {
            launchTemplateName = launchTemplateNameVal
        }

    val groupRequest =
        UpdateAutoScalingGroupRequest {
            maxSize = 3
            serviceLinkedRoleArn = serviceLinkedRoleARNVal
            autoScalingGroupName = groupName
            launchTemplate = templateSpecification
        }

    val groupsRequestWaiter =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.updateAutoScalingGroup(groupRequest)
        autoScalingClient.waitUntilGroupExists(groupsRequestWaiter)
        println("You successfully updated the Auto Scaling group  $groupName")
    }
}

suspend fun createAutoScalingGroup(
```

```
    groupName: String,
    launchTemplateNameVal: String,
    serviceLinkedRoleARNVal: String,
    vpcZoneIdVal: String,
) {
    val templateSpecification =
        LaunchTemplateSpecification {
            launchTemplateName = launchTemplateNameVal
        }

    val request =
        CreateAutoScalingGroupRequest {
            autoScalingGroupName = groupName
            availabilityZones = listOf("us-east-1a")
            launchTemplate = templateSpecification
            maxSize = 1
            minSize = 1
            vpcZoneIdentifier = vpcZoneIdVal
            serviceLinkedRoleArn = serviceLinkedRoleARNVal
        }

    // This object is required for the waiter call.
    val groupsRequestWaiter =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.createAutoScalingGroup(request)
        autoScalingClient.waitUntilGroupExists(groupsRequestWaiter)
        println("$groupName was created!")
    }
}

suspend fun describeAutoScalingInstance(id: String) {
    val describeAutoScalingInstancesRequest =
        DescribeAutoScalingInstancesRequest {
            instanceIds = listOf(id)
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        val response =
 autoScalingClient.describeAutoScalingInstances(describeAutoScalingInstancesRequest)
        response.autoScalingInstances?.forEach { group ->
```

```
            println("The instance lifecycle state is: ${group.lifecycleState}")
        }
    }
}

suspend fun enableMetricsCollection(groupName: String?) {
    val collectionRequest =
        EnableMetricsCollectionRequest {
            autoScalingGroupName = groupName
            metrics = listOf("GroupMaxSize")
            granularity = "1Minute"
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.enableMetricsCollection(collectionRequest)
        println("The enable metrics collection operation was successful")
    }
}

suspend fun getSpecificAutoScaling(groupName: String): String {
    var instanceId = ""
    val scalingGroupsRequest =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        val response =
 autoScalingClient.describeAutoScalingGroups(scalingGroupsRequest)
        response.autoScalingGroups?.forEach { group ->
            println("The group name is ${group.autoScalingGroupName}")
            println("The group ARN is ${group.autoScalingGroupArn}")

            group.instances?.forEach { instance ->
                instanceId = instance.instanceId.toString()
            }
        }
    }
    return instanceId
}

suspend fun describeAccountLimits() {
    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
```

```
        val response =
 autoScalingClient.describeAccountLimits(DescribeAccountLimitsRequest {})
        println("The max number of Auto Scaling groups is
 ${response.maxNumberOfAutoScalingGroups}")
        println("The current number of Auto Scaling groups is
 ${response.numberOfAutoScalingGroups}")
    }
}

suspend fun terminateInstanceInAutoScalingGroup(instanceIdVal: String) {
    val request =
        TerminateInstanceInAutoScalingGroupRequest {
            instanceId = instanceIdVal
            shouldDecrementDesiredCapacity = false
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.terminateInstanceInAutoScalingGroup(request)
        println("You have terminated instance $instanceIdVal")
    }
}

suspend fun deleteSpecificAutoScalingGroup(groupName: String) {
    val deleteAutoScalingGroupRequest =
        DeleteAutoScalingGroupRequest {
            autoScalingGroupName = groupName
            forceDelete = true
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.deleteAutoScalingGroup(deleteAutoScalingGroupRequest)
        println("You successfully deleted $groupName")
    }
}
```

- Consulte detalhes da API nos tópicos a seguir na Referência de API do AWS SDK para Kotlin.
  - CreateAutoScalingGroup
  - DeleteAutoScalingGroup
  - DescribeAutoScalingGroups
  - DescribeAutoScalingInstances

- DescribeScalingActivities
- DisableMetricsCollection
- EnableMetricsCollection
- SetDesiredCapacity
- TerminateInstanceInAutoScalingGroup
- UpdateAutoScalingGroup

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
namespace AutoScaling;

use Aws\AutoScaling\AutoScalingClient;
use Aws\CloudWatch\CloudWatchClient;
use Aws\Ec2\Ec2Client;
use AwsUtilities\AWSServiceClass;
use AwsUtilities\RunnableExample;

class GettingStartedWithAutoScaling implements RunnableExample
{
    protected Ec2Client $ec2Client;
    protected AutoScalingClient $autoScalingClient;
    protected AutoScalingService $autoScalingService;
    protected CloudWatchClient $cloudWatchClient;
    protected string $templateName;
    protected string $autoScalingGroupName;
    protected array $role;

    public function runExample()
    {
        echo("\n");
        echo("--------------------------------------\n");
```

```
        print("Welcome to the Amazon EC2 Auto Scaling getting started demo using
 PHP!\n");
        echo("-------------------------------------\n");

        $clientArgs = [
            'region' => 'us-west-2',
            'version' => 'latest',
            'profile' => 'default',
        ];
        $uniqid = uniqid();

        $this->autoScalingClient = new AutoScalingClient($clientArgs);
        $this->autoScalingService = new AutoScalingService($this-
>autoScalingClient);
        $this->cloudWatchClient = new CloudWatchClient($clientArgs);

        AWSServiceClass::$waitTime = 5;
        AWSServiceClass::$maxWaitAttempts = 20;

        /**
         * Step 0: Create an EC2 launch template that you'll use to create an
 Auto Scaling group.
         */
        $this->ec2Client = new EC2Client($clientArgs);
        $this->templateName = "example_launch_template_$uniqid";
        $instanceType = "t1.micro";
        $amiId = "ami-0ca285d4c2cda3300";
        $launchTemplate = $this->ec2Client->createLaunchTemplate(
            [
            'LaunchTemplateName' => $this->templateName,
            'LaunchTemplateData' => [
                'InstanceType' => $instanceType,
                'ImageId' => $amiId,
            ]
            ]
        );

        /**
         * Step 1: CreateAutoScalingGroup: pass it the launch template you
 created in step 0.
         */
        $availabilityZones[] = $this->ec2Client->describeAvailabilityZones([])
['AvailabilityZones'][1]['ZoneName'];
```

```
        $this->autoScalingGroupName = "demoAutoScalingGroupName_$uniqid";
        $minSize = 1;
        $maxSize = 1;
        $launchTemplateId = $launchTemplate['LaunchTemplate']
['LaunchTemplateId'];
        $this->autoScalingService->createAutoScalingGroup(
            $this->autoScalingGroupName,
            $availabilityZones,
            $minSize,
            $maxSize,
            $launchTemplateId
        );

        $this->autoScalingService->waitUntilGroupInService([$this-
>autoScalingGroupName]);
        $autoScalingGroup = $this->autoScalingService-
>describeAutoScalingGroups([$this->autoScalingGroupName]);

        /**
         * Step 2: DescribeAutoScalingInstances: show that one instance has
 launched.
         */
        $instanceIds = [$autoScalingGroup['AutoScalingGroups'][0]['Instances'][0]
['InstanceId']];
        $instances = $this->autoScalingService-
>describeAutoScalingInstances($instanceIds);
        echo "The Auto Scaling group {$this->autoScalingGroupName} was created
 successfully.\n";
        echo count($instances['AutoScalingInstances']) . " instances were created
 for the group.\n";
        echo $autoScalingGroup['AutoScalingGroups'][0]['MaxSize'] . " is the max
 number of instances for the group.\n";

        /**
         * Step 3: EnableMetricsCollection: enable all metrics or a subset.
         */
        $this->autoScalingService->enableMetricsCollection($this-
>autoScalingGroupName, "1Minute");

        /**
         * Step 4: UpdateAutoScalingGroup: update max size to 3.
         */
        echo "Updating the max number of instances to 3.\n";
```

```
        $this->autoScalingService->updateAutoScalingGroup($this-
>autoScalingGroupName, ['MaxSize' => 3]);

        /**
         * Step 5: DescribeAutoScalingGroups: show the current state of the
 group.
         */
        $autoScalingGroup = $this->autoScalingService-
>describeAutoScalingGroups([$this->autoScalingGroupName]);
        echo $autoScalingGroup['AutoScalingGroups'][0]['MaxSize'];
        echo " is the updated max number of instances for the group.\n";

        $limits = $this->autoScalingService->describeAccountLimits();
        echo "Here are your account limits:\n";
        echo "MaxNumberOfAutoScalingGroups:
 {$limits['MaxNumberOfAutoScalingGroups']}\n";
        echo "MaxNumberOfLaunchConfigurations:
 {$limits['MaxNumberOfLaunchConfigurations']}\n";
        echo "NumberOfAutoScalingGroups:
 {$limits['NumberOfAutoScalingGroups']}\n";
        echo "NumberOfLaunchConfigurations:
 {$limits['NumberOfLaunchConfigurations']}\n";

        /**
         * Step 6: SetDesiredCapacity: set desired capacity to 2.
         */
        $this->autoScalingService->setDesiredCapacity($this-
>autoScalingGroupName, 2);
        sleep(10); // Wait for the group to start processing the request.
        $this->autoScalingService->waitUntilGroupInService([$this-
>autoScalingGroupName]);

        /**
         * Step 7: DescribeAutoScalingInstances: show that two instances are
 launched.
         */
        $autoScalingGroups = $this->autoScalingService-
>describeAutoScalingGroups([$this->autoScalingGroupName]);
        foreach ($autoScalingGroups['AutoScalingGroups'] as $autoScalingGroup) {
            echo "There is a group named:
 {$autoScalingGroup['AutoScalingGroupName']}";
            echo "with an ARN of {$autoScalingGroup['AutoScalingGroupARN']}.\n";
            foreach ($autoScalingGroup['Instances'] as $instance) {
```

```
                echo "{$autoScalingGroup['AutoScalingGroupName']} has an instance
 with id of: ";
                echo "{$instance['InstanceId']} and a lifecycle state of:
 {$instance['LifecycleState']}.\n";
            }
        }

        /**
         * Step 8: TerminateInstanceInAutoScalingGroup: terminate one of the
 instances in the group.
         */
        $this->autoScalingService-
>terminateInstanceInAutoScalingGroup($instance['InstanceId'], false);
        do {
            sleep(10);
            $instances = $this->autoScalingService-
>describeAutoScalingInstances([$instance['InstanceId']]);
        } while (count($instances['AutoScalingInstances']) > 0);
        do {
            sleep(10);
            $autoScalingGroups = $this->autoScalingService-
>describeAutoScalingGroups([$this->autoScalingGroupName]);
            $instances = $autoScalingGroups['AutoScalingGroups'][0]['Instances'];
        } while (count($instances) < 2);
        $this->autoScalingService->waitUntilGroupInService([$this-
>autoScalingGroupName]);
        foreach ($autoScalingGroups['AutoScalingGroups'] as $autoScalingGroup) {
            echo "There is a group named:
 {$autoScalingGroup['AutoScalingGroupName']}";
            echo "with an ARN of {$autoScalingGroup['AutoScalingGroupARN']}.\n";
            foreach ($autoScalingGroup['Instances'] as $instance) {
                echo "{$autoScalingGroup['AutoScalingGroupName']} has an instance
 with id of: ";
                echo "{$instance['InstanceId']} and a lifecycle state of:
 {$instance['LifecycleState']}.\n";
            }
        }

        /**
         * Step 9: DescribeScalingActivities: list the scaling activities that
 have occurred for the group so far.
         */
        $activities = $this->autoScalingService-
>describeScalingActivities($autoScalingGroup['AutoScalingGroupName']);
```

```
        echo "We found " . count($activities['Activities']) . " activities.\n";
        foreach ($activities['Activities'] as $activity) {
            echo "{$activity['ActivityId']} - {$activity['StartTime']} -
{$activity['Description']}\n";
        }

        /**
         * Step 10: Use the Amazon CloudWatch API to get and show some metrics
collected for the group.
         */
        $metricsNamespace = 'AWS/AutoScaling';
        $metricsDimensions = [
            [
                'Name' => 'AutoScalingGroupName',
                'Value' => $autoScalingGroup['AutoScalingGroupName'],
            ],
        ];
        $metrics = $this->cloudWatchClient->listMetrics(
            [
            'Dimensions' => $metricsDimensions,
            'Namespace' => $metricsNamespace,
            ]
        );
        foreach ($metrics['Metrics'] as $metric) {
            $timespan = 5;
            if ($metric['MetricName'] != 'GroupTotalCapacity' &&
$metric['MetricName'] != 'GroupMaxSize') {
                continue;
            }
            echo "Over the last $timespan minutes, {$metric['MetricName']}
recorded:\n";
            $stats = $this->cloudWatchClient->getMetricStatistics(
                [
                'Dimensions' => $metricsDimensions,
                'EndTime' => time(),
                'StartTime' => time() - (5 * 60),
                'MetricName' => $metric['MetricName'],
                'Namespace' => $metricsNamespace,
                'Period' => 60,
                'Statistics' => ['Sum'],
                ]
            );
            foreach ($stats['Datapoints'] as $stat) {
                echo "{$stat['Timestamp']}: {$stat['Sum']}\n";
```

```
            }
        }

        return $instances;
    }

    public function cleanUp()
    {
        /**
         * Step 11: DisableMetricsCollection: disable all metrics.
         */
        $this->autoScalingService->disableMetricsCollection($this-
>autoScalingGroupName);

        /**
         * Step 12: DeleteAutoScalingGroup: to delete the group you must stop all
 instances.
         * - UpdateAutoScalingGroup with MinSize=0
         * - TerminateInstanceInAutoScalingGroup for each instance,
         *     specify ShouldDecrementDesiredCapacity=True. Wait for instances to
 stop.
         * - Now you can delete the group.
         */
        $this->autoScalingService->updateAutoScalingGroup($this-
>autoScalingGroupName, ['MinSize' => 0]);
        $this->autoScalingService->terminateAllInstancesInAutoScalingGroup($this-
>autoScalingGroupName);
        $this->autoScalingService->waitUntilGroupInService([$this-
>autoScalingGroupName]);
        $this->autoScalingService->deleteAutoScalingGroup($this-
>autoScalingGroupName);

        /**
         * Step 13: Delete launch template.
         */
        $this->ec2Client->deleteLaunchTemplate(
            [
            'LaunchTemplateName' => $this->templateName,
            ]
        );
    }

    public function helloService()
    {
```

```
        $autoScalingClient = new AutoScalingClient([
            'region' => 'us-west-2',
            'version' => 'latest',
            'profile' => 'default',
        ]);

        $groups = $autoScalingClient->describeAutoScalingGroups([]);
        var_dump($groups);
    }
}
```

- Para obter detalhes da API, consulte os tópicos a seguir na Referência da API AWS SDK para PHP .
  - CreateAutoScalingGroup
  - DeleteAutoScalingGroup
  - DescribeAutoScalingGroups
  - DescribeAutoScalingInstances
  - DescribeScalingActivities
  - DisableMetricsCollection
  - EnableMetricsCollection
  - SetDesiredCapacity
  - TerminateInstanceInAutoScalingGroup
  - UpdateAutoScalingGroup

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

Execute um cenário interativo em um prompt de comando.

```
def run_scenario(as_wrapper: AutoScalingWrapper, svc_helper: ServiceHelper) ->
 None:
    """
    Runs the scenario demonstrating the management of Auto Scaling groups and
 instances.

    :param as_wrapper: An instance of the AutoScalingWrapper that manages Auto
 Scaling groups.
    :param svc_helper: An instance of the ServiceHelper that interacts with AWS
 services.
    :return: None
    """
    logging.basicConfig(level=logging.INFO, format="%(levelname)s: %(message)s")

    logger.info("Starting the Amazon EC2 Auto Scaling demo.")

    print("-" * 88)
    print(
        "Welcome to the Amazon EC2 Auto Scaling demo for managing groups and
 instances."
    )
    print("-" * 88)

    print(
        "This example requires a launch template that specifies how to create "
        "EC2 instances. You can use an existing template or create a new one."
    )
    template_name = q.ask(
        "Enter the name of an existing launch template or press Enter to create a
 new one: "
    )
    template = None
    if template_name:
        template = svc_helper.get_template(template_name)
    if template is None:
        inst_type = "t1.micro"
        ami_id = "ami-0ca285d4c2cda3300"
        print("Let's create a launch template with the following
 specifications:")
        print(f"\tInstanceType: {inst_type}")
        print(f"\tAMI ID: {ami_id}")
        template_name = q.ask("Enter a name for the template: ", q.non_empty)
        template = svc_helper.create_template(template_name, inst_type, ami_id)
```

```
    print("-" * 88)

    print("Let's create an Auto Scaling group.")
    group_name = q.ask("Enter a name for the group: ", q.non_empty)
    zones = svc_helper.get_availability_zones()
    print("EC2 instances can be created in the following Availability Zones:")
    for index, zone in enumerate(zones):
        print(f"\t{index+1}. {zone}")
    print(f"\t{len(zones)+1}. All zones")
    zone_sel = q.ask(
        "Which zone do you want to use? ", q.is_int, q.in_range(1, len(zones) +
1)
    )
    group_zones = [zones[zone_sel - 1]] if zone_sel <= len(zones) else zones
    print(f"Creating group {group_name}...")
    as_wrapper.create_autoscaling_group(group_name, group_zones, template_name,
1, 1)
    wait(10)
    group = as_wrapper.describe_group(group_name)
    logger.info("Created Auto Scaling group %s.", group_name)
    print("Created group:")
    pp(group)
    print("Waiting for instance to start...")
    wait_for_group(group_name, as_wrapper)
    print("-" * 88)

    use_metrics = q.ask(
        "Do you want to collect metrics about Amazon EC2 Auto Scaling during this
demo (y/n)? ",
        q.is_yesno,
    )
    if use_metrics:
        as_wrapper.enable_metrics(
            group_name,
            [
                "GroupMinSize",
                "GroupMaxSize",
                "GroupDesiredCapacity",
                "GroupInServiceInstances",
                "GroupTotalInstances",
            ],
        )
        logger.info("Enabled metrics for Auto Scaling group %s.", group_name)
        print(f"Metrics enabled for {group_name}.")
```

```
    print("-" * 88)

    print(f"Let's update the maximum number of instances in {group_name} from 1
to 3.")
    q.ask("Press Enter when you're ready.")
    as_wrapper.update_group(group_name, MaxSize=3)
    group = as_wrapper.describe_group(group_name)
    logger.info("Updated maximum size for group %s to 3.", group_name)
    print("The group still has one running instance, but can have up to three:")
    print_simplified_group(group)
    print("-" * 88)

    print(f"Let's update the desired capacity of {group_name} from 1 to 2.")
    q.ask("Press Enter when you're ready.")
    as_wrapper.set_desired_capacity(group_name, 2)
    wait(10)
    group = as_wrapper.describe_group(group_name)
    logger.info("Set desired capacity for group %s to 2.", group_name)
    print("Here's the current state of the group:")
    print_simplified_group(group)
    print("-" * 88)
    print("Waiting for the new instance to start...")
    instance_ids = wait_for_group(group_name, as_wrapper)
    print("-" * 88)

    print(f"Let's terminate one of the instances in {group_name}.")
    print("Because the desired capacity is 2, another instance will start.")
    print("The currently running instances are:")
    for index, inst_id in enumerate(instance_ids):
        print(f"\t{index+1}. {inst_id}")
    inst_sel = q.ask(
        "Which instance do you want to stop? ",
        q.is_int,
        q.in_range(1, len(instance_ids) + 1),
    )
    print(f"Stopping {instance_ids[inst_sel-1]}...")
    as_wrapper.terminate_instance(instance_ids[inst_sel - 1], False)
    wait(10)
    group = as_wrapper.describe_group(group_name)
    logger.info(
        "Terminated instance %s in group %s.", instance_ids[inst_sel - 1],
group_name
    )
    print(f"Here's the state of {group_name}:")
```

```
    print_simplified_group(group)
    print("Waiting for the scaling activities to complete...")
    wait_for_group(group_name, as_wrapper)
    print("-" * 88)

    print(f"Let's get a report of scaling activities for {group_name}.")
    q.ask("Press Enter when you're ready.")
    activities = as_wrapper.describe_scaling_activities(group_name)
    logger.info(
        "Retrieved %d scaling activities for group %s.", len(activities),
group_name
    )
    print(
        f"Found {len(activities)} activities.\n"
        f"Activities are ordered with the most recent one first:"
    )
    for act in activities:
        pp(act)
    print("-" * 88)

    if use_metrics:
        print("Let's look at CloudWatch metrics.")
        metric_namespace = "AWS/AutoScaling"
        metric_dimensions = [{"Name": "AutoScalingGroupName", "Value":
group_name}]
        print(f"The following metrics are enabled for {group_name}:")
        done = False
        while not done:
            metrics = svc_helper.get_metrics(metric_namespace, metric_dimensions)
            for index, metric in enumerate(metrics):
                print(f"\t{index+1}. {metric.name}")
            print(f"\t{len(metrics)+1}. None")
            metric_sel = q.ask(
                "Which metric do you want to see? ",
                q.is_int,
                q.in_range(1, len(metrics) + 1),
            )
            if metric_sel < len(metrics) + 1:
                span = 5
                metric = metrics[metric_sel - 1]
                print(f"Over the last {span} minutes, {metric.name} recorded:")
                # CloudWatch metric times are in the UTC+0 time zone.
                now = datetime.now(timezone.utc)
                metric_data = svc_helper.get_metric_statistics(
```

```
                    metric_dimensions, metric, now - timedelta(minutes=span), now
                )
                pp(metric_data)
                if not q.ask("Do you want to see another metric (y/n)? ",
q.is_yesno):
                    done = True
            else:
                done = True

    print(f"Let's clean up.")
    q.ask("Press Enter when you're ready.")
    if use_metrics:
        print(f"Stopping metrics collection for {group_name}.")
        as_wrapper.disable_metrics(group_name)
        logger.info("Disabled metrics collection for group %s.", group_name)

    print(
        "You must terminate all instances in the group before you can delete the
group."
    )
    print("Set minimum size to 0.")
    as_wrapper.update_group(group_name, MinSize=0)
    group = as_wrapper.describe_group(group_name)
    instance_ids = [inst["InstanceId"] for inst in group["Instances"]]
    for inst_id in instance_ids:
        print(f"Stopping {inst_id}.")
        as_wrapper.terminate_instance(inst_id, True)
        logger.info("Terminated instance %s in group %s.", inst_id, group_name)
    print("Waiting for instances to stop...")
    wait_for_instances(instance_ids, as_wrapper)
    print(f"Deleting {group_name}.")
    as_wrapper.delete_autoscaling_group(group_name)
    logger.info("Deleted Auto Scaling group %s.", group_name)
    print("-" * 88)

    if template is not None:
        if q.ask(
            f"Do you want to delete launch template {template_name} used in this
demo (y/n)? "
        ):
            svc_helper.delete_template(template_name)
            logger.info("Deleted launch template %s.", template_name)
            print("Template deleted.")
```

```
    print("\nThanks for watching!")
    print("-" * 88)


if __name__ == "__main__":
    try:
        wrapper = AutoScalingWrapper(boto3.client("autoscaling"))
        helper = ServiceHelper(boto3.client("ec2"), boto3.resource("cloudwatch"))
        run_scenario(wrapper, helper)
    except Exception:
        logger.exception("Something went wrong with the demo!")
```

Defina as funções que são chamadas pelo cenário para gerenciar modelos e métricas de lançamento. Essas funções envolvem o Amazon EC2 e CloudWatch as ações.

```
class ServiceHelper:
    """Encapsulates Amazon EC2 and CloudWatch actions for the example."""

    def __init__(self, ec2_client, cloudwatch_resource):
        """
        :param ec2_client: A Boto3 Amazon EC2 client.
        :param cloudwatch_resource: A Boto3 CloudWatch resource.
        """
        self.ec2_client = ec2_client
        self.cloudwatch_resource = cloudwatch_resource

    def get_template(self, template_name: str) -> dict:
        """
        Gets a launch template. Launch templates specify configuration for
 instances
        that are launched by Amazon EC2 Auto Scaling.

        :param template_name: The name of the template to look up.
        :return: The template, if it exists.
        :raises ClientError: If there is an error retrieving the launch template.
        """
        try:
            response = self.ec2_client.describe_launch_templates(
                LaunchTemplateNames=[template_name]
            )
            template = response["LaunchTemplates"][0]
```

```
            logger.info("Launch template %s retrieved successfully.",
template_name)
            return template
        except ClientError as err:
            if (
                err.response["Error"]["Code"]
                == "InvalidLaunchTemplateName.NotFoundException"
            ):
                logger.warning("Launch template %s does not exist.",
template_name)
            else:
                logger.error(
                    "Couldn't verify launch template %s. Error: %s: %s",
                    template_name,
                    err.response["Error"]["Code"],
                    err.response["Error"]["Message"],
                )
                raise

    def create_template(self, template_name: str, inst_type: str, ami_id: str) ->
dict:
        """
        Creates an Amazon EC2 launch template to use with Amazon EC2 Auto
Scaling.

        :param template_name: The name to give to the template.
        :param inst_type: The type of the instance, such as t1.micro.
        :param ami_id: The ID of the Amazon Machine Image (AMI) to use when
creating
                       an instance.
        :return: Information about the newly created template.
        :raises ClientError: If there is an error creating the launch template.
        """
        try:
            response = self.ec2_client.create_launch_template(
                LaunchTemplateName=template_name,
                LaunchTemplateData={"InstanceType": inst_type, "ImageId":
ami_id},
            )
            template = response["LaunchTemplate"]
            logger.info(
                "Created launch template %s with instance type %s and AMI ID
%s.",
                template_name,
```

```
                inst_type,
                ami_id,
            )
            return template
        except ClientError as err:
            logger.error(
                "Couldn't create launch template %s. Error: %s: %s",
                template_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise

    def delete_template(self, template_name: str) -> None:
        """
        Deletes a launch template.

        :param template_name: The name of the template to delete.
        :raises ClientError: If there is an error deleting the launch template.
        """
        try:

self.ec2_client.delete_launch_template(LaunchTemplateName=template_name)
            logger.info("Deleted launch template %s.", template_name)
        except ClientError as err:
            logger.error(
                "Couldn't delete launch template %s. Error: %s: %s",
                template_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise

    def get_availability_zones(self) -> list:
        """
        Gets a list of Availability Zones in the AWS Region of the Amazon EC2
client.

        :return: The list of Availability Zones for the client Region.
        :raises ClientError: If there is an error retrieving availability zones.
        """
        try:
            response = self.ec2_client.describe_availability_zones()
            zones = [zone["ZoneName"] for zone in response["AvailabilityZones"]]
```

```
            logger.info("Retrieved availability zones: %s.", ", ".join(zones))
            return zones
        except ClientError as err:
            logger.error(
                "Couldn't get availability zones. Error: %s: %s",
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise

    def get_metrics(self, namespace: str, dimensions: list) -> list:
        """
        Gets a list of CloudWatch metrics filtered by namespace and dimensions.

        :param namespace: The namespace of the metrics to look up.
        :param dimensions: The dimensions of the metrics to look up.
        :return: The list of metrics.
        :raises ClientError: If there is an error retrieving CloudWatch metrics.
        """
        try:
            metrics = list(
                self.cloudwatch_resource.metrics.filter(
                    Namespace=namespace, Dimensions=dimensions
                )
            )
            logger.info(
                "Retrieved metrics for namespace %s with dimensions %s.",
                namespace,
                dimensions,
            )
            return metrics
        except ClientError as err:
            logger.error(
                "Couldn't get metrics for %s, %s. Error: %s: %s",
                namespace,
                dimensions,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise

    @staticmethod
    def get_metric_statistics(
        dimensions: list, metric, start: datetime, end: datetime
```

```
    ) -> list:
        """
        Gets statistics for a CloudWatch metric within a specified time span.

        :param dimensions: The dimensions of the metric.
        :param metric: The metric to look up.
        :param start: The start of the time span for retrieved metrics.
        :param end: The end of the time span for retrieved metrics.
        :return: The list of data points found for the specified metric.
        :raises ClientError: If there is an error retrieving metric statistics.
        """
        try:
            response = metric.get_statistics(
                Dimensions=dimensions,
                StartTime=start,
                EndTime=end,
                Period=60,
                Statistics=["Sum"],
            )
            data = response["Datapoints"]
            logger.info("Retrieved statistics for metric %s.", metric.name)
            return data
        except ClientError as err:
            logger.error(
                "Couldn't get statistics for metric %s. Error: %s: %s",
                metric.name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise


def print_simplified_group(group: dict) -> None:
    """
    Prints a subset of data for an Auto Scaling group.

    :param group: The Auto Scaling group data to print.
    :return: None
    """
    print(group["AutoScalingGroupName"])
    print(f"\tLaunch template: {group['LaunchTemplate']['LaunchTemplateName']}")
    print(
        f"\tMin: {group['MinSize']}, Max: {group['MaxSize']}, Desired:
 {group['DesiredCapacity']}"
```

```
    )
    if group["Instances"]:
        print(f"\tInstances:")
        for inst in group["Instances"]:
            print(f"\t\t{inst['InstanceId']}: {inst['LifecycleState']}")


def wait_for_group(group_name: str, as_wrapper: AutoScalingWrapper) -> list:
    """
    Waits for instances to start or stop in an Auto Scaling group.
    Prints the data for each instance after scaling activities are complete.

    :param group_name: The name of the Auto Scaling group.
    :param as_wrapper: The AutoScalingWrapper that manages Auto Scaling groups.
    :return: A list of instance IDs in the group.
    """
    group = as_wrapper.describe_group(group_name)
    instance_ids = [i["InstanceId"] for i in group["Instances"]]
    return wait_for_instances(instance_ids, as_wrapper)


def wait_for_instances(instance_ids: list, as_wrapper: AutoScalingWrapper) ->
 list:
    """
    Waits for instances to start or stop in an Auto Scaling group.
    Prints the data for each instance after scaling activities are complete.

    :param instance_ids: A list of instance IDs to wait for.
    :param as_wrapper: The AutoScalingWrapper that manages Auto Scaling groups.
    :return: A list of instance IDs that were waited on.
    """
    ready = False
    instances = []
    while not ready:
        instances = as_wrapper.describe_instances(instance_ids) if instance_ids
 else []
        if all([x["LifecycleState"] in ["Terminated", "InService"] for x in
 instances]):
            ready = True
        else:
            wait(10)
    if instances:
        print(
```

```
            f"Here are the details of the instance{'s' if len(instances) > 1 else
''}:"
        )
        for instance in instances:
            pp(instance)
    return instance_ids
```

- Para ver detalhes da API, consulte os tópicos a seguir na Referência da API do SDK da AWS para Python (Boto3).
  - CreateAutoScalingGroup
  - DeleteAutoScalingGroup
  - DescribeAutoScalingGroups
  - DescribeAutoScalingInstances
  - DescribeScalingActivities
  - DisableMetricsCollection
  - EnableMetricsCollection
  - SetDesiredCapacity
  - TerminateInstanceInAutoScalingGroup
  - UpdateAutoScalingGroup

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
[package]
name = "autoscaling-code-examples"
version = "0.1.0"
```

```
authors = ["Doug Schwartz <dougsch@amazon.com>", "David Souther
 <dpsouth@amazon.com>"]
edition = "2021"

# See more keys and their definitions at https://doc.rust-lang.org/cargo/
reference/manifest.html

[dependencies]
aws-config = { version = "1.0.1", features = ["behavior-version-latest"] }
aws-sdk-autoscaling = { version = "1.3.0" }
aws-sdk-ec2 = { version = "1.3.0" }
aws-types = { version = "1.0.1" }
tokio = { version = "1.20.1", features = ["full"] }
clap = { version = "4.4", features = ["derive"] }
tracing-subscriber = { version = "0.3.15", features = ["env-filter"] }
anyhow = "1.0.75"
tracing = "0.1.37"
tokio-stream = "0.1.14"


use std::{collections::BTreeSet, fmt::Display};

use anyhow::anyhow;
use autoscaling_code_examples::scenario::{AutoScalingScenario, ScenarioError};
use tracing::{info, warn};

async fn show_scenario_description(scenario: &AutoScalingScenario, event: &str) {
    let description = scenario.describe_scenario().await;
    info!("DescribeAutoScalingInstances: {event}\n{description}");
}

#[derive(Default, Debug)]
struct Warnings(Vec<String>);

impl Warnings {
    pub fn push(&mut self, warning: &str, error: ScenarioError) {
        let formatted = format!("{warning}: {error}");
        warn!("{formatted}");
        self.0.push(formatted);
    }

    pub fn is_empty(&self) -> bool {
        self.0.is_empty()
    }
```

```
}

impl Display for Warnings {
    fn fmt(&self, f: &mut std::fmt::Formatter<'_>) -> std::fmt::Result {
        writeln!(f, "Warnings:")?;
        for warning in &self.0 {
            writeln!(f, "{: >4}- {warning}", "")?;
        }
        Ok(())
    }
}

#[tokio::main]
async fn main() -> Result<(), anyhow::Error> {
    tracing_subscriber::fmt::init();

    let shared_config = aws_config::from_env().load().await;

    let mut warnings = Warnings::default();

    // 1. Create an EC2 launch template that you'll use to create an auto scaling
 group. Bonus: use SDK with EC2.CreateLaunchTemplate to create the launch
 template.
    // 2. CreateAutoScalingGroup: pass it the launch template you created in step
 0. Give it min/max of 1 instance.
    // 4. EnableMetricsCollection: enable all metrics or a subset.
    let scenario = match
 AutoScalingScenario::prepare_scenario(&shared_config).await {
        Ok(scenario) => scenario,
        Err(errs) => {
            let err_str = errs
                .into_iter()
                .map(|e| e.to_string())
                .collect::<Vec<String>>()
                .join(", ");
            return Err(anyhow!("Failed to initialize scenario: {err_str}"));
        }
    };

    info!("Prepared autoscaling scenario:\n{scenario}");

    let stable = scenario.wait_for_stable(1).await;
    if let Err(err) = stable {
        warnings.push(
```

```
            "There was a problem while waiting for group to be stable",
            err,
        );
    }

    // 3. DescribeAutoScalingInstances: show that one instance has launched.
    show_scenario_description(
        &scenario,
        "show that the group was created and one instance has launched",
    )
    .await;

    // 5. UpdateAutoScalingGroup: update max size to 3.
    let scale_max_size = scenario.scale_max_size(3).await;
    if let Err(err) = scale_max_size {
        warnings.push("There was a problem scaling max size", err);
    }

    // 6. DescribeAutoScalingGroups: the current state of the group
    show_scenario_description(
        &scenario,
        "show the current state of the group after setting max size",
    )
    .await;

    // 7. SetDesiredCapacity: set desired capacity to 2.
    let scale_desired_capacity = scenario.scale_desired_capacity(2).await;
    if let Err(err) = scale_desired_capacity {
        warnings.push("There was a problem setting desired capacity", err);
    }

    //   Wait for a second instance to launch.
    let stable = scenario.wait_for_stable(2).await;
    if let Err(err) = stable {
        warnings.push(
            "There was a problem while waiting for group to be stable",
            err,
        );
    }

    // 8. DescribeAutoScalingInstances: show that two instances are launched.
    show_scenario_description(
        &scenario,
        "show that two instances are launched after setting desired capacity",
```

```
    )
    .await;

    let ids_before = scenario
        .list_instances()
        .await
        .map(|v| v.into_iter().collect::<BTreeSet<_>>())
        .unwrap_or_default();

    // 9. TerminateInstanceInAutoScalingGroup: terminate one of the instances in
the group.
    let terminate_some_instance = scenario.terminate_some_instance().await;
    if let Err(err) = terminate_some_instance {
        warnings.push("There was a problem replacing an instance", err);
    }

    let wait_after_terminate = scenario.wait_for_stable(1).await;
    if let Err(err) = wait_after_terminate {
        warnings.push(
            "There was a problem waiting after terminating an instance",
            err,
        );
    }

    let wait_scale_up_after_terminate = scenario.wait_for_stable(2).await;
    if let Err(err) = wait_scale_up_after_terminate {
        warnings.push(
            "There was a problem waiting for scale up after terminating an
instance",
            err,
        );
    }

    let ids_after = scenario
        .list_instances()
        .await
        .map(|v| v.into_iter().collect::<BTreeSet<_>>())
        .unwrap_or_default();

    let difference = ids_after.intersection(&ids_before).count();
    if !(difference == 1 && ids_before.len() == 2 && ids_after.len() == 2) {
        warnings.push(
            "Before and after set not different",
            ScenarioError::with(format!("{difference}")),
```

```
        );
    }

    // 10. DescribeScalingActivities: list the scaling activities that have
 occurred for the group so far.
    show_scenario_description(
        &scenario,
        "list the scaling activities that have occurred for the group so far",
    )
    .await;

    // 11. DisableMetricsCollection
    let scale_group = scenario.scale_group_to_zero().await;
    if let Err(err) = scale_group {
        warnings.push("There was a problem scaling the group to 0", err);
    }
    show_scenario_description(&scenario, "Scenario scaled to 0").await;

    // 12. DeleteAutoScalingGroup (to delete the group you must stop all
 instances):
    // 13. Delete LaunchTemplate.
    let clean_scenario = scenario.clean_scenario().await;
    if let Err(errs) = clean_scenario {
        for err in errs {
            warnings.push("There was a problem cleaning the scenario", err);
        }
    } else {
        info!("The scenario has been cleaned up!");
    }

    if warnings.is_empty() {
        Ok(())
    } else {
        Err(anyhow!(
            "There were warnings during scenario execution:\n{warnings}"
        ))
    }
}

pub mod scenario;


use std::{
    error::Error,
```

```
    fmt::{Debug, Display},
    time::{Duration, SystemTime},
};

use anyhow::anyhow;
use aws_config::SdkConfig;
use aws_sdk_autoscaling::{
    error::{DisplayErrorContext, ProvideErrorMetadata},
    types::{Activity, AutoScalingGroup, LaunchTemplateSpecification},
};
use aws_sdk_ec2::types::RequestLaunchTemplateData;
use tracing::trace;

const LAUNCH_TEMPLATE_NAME: &str =
 "SDK_Code_Examples_EC2_Autoscaling_template_from_Rust_SDK";
const AUTOSCALING_GROUP_NAME: &str =
 "SDK_Code_Examples_EC2_Autoscaling_Group_from_Rust_SDK";
const MAX_WAIT: Duration = Duration::from_secs(5 * 60); // Wait at most 25
 seconds.
const WAIT_TIME: Duration = Duration::from_millis(500); // Wait half a second at
 a time.

struct Waiter {
    start: SystemTime,
    max: Duration,
}

impl Waiter {
    fn new() -> Self {
        Waiter {
            start: SystemTime::now(),
            max: MAX_WAIT,
        }
    }

    async fn sleep(&self) -> Result<(), ScenarioError> {
        if SystemTime::now()
            .duration_since(self.start)
            .unwrap_or(Duration::MAX)
            > self.max
        {
            Err(ScenarioError::with(
                "Exceeded maximum wait duration for stable group",
            ))
```

```
        } else {
            tokio::time::sleep(WAIT_TIME).await;
            Ok(())
        }
    }
}

pub struct AutoScalingScenario {
    ec2: aws_sdk_ec2::Client,
    autoscaling: aws_sdk_autoscaling::Client,
    launch_template_arn: String,
    auto_scaling_group_name: String,
}

impl Display for AutoScalingScenario {
    fn fmt(&self, f: &mut std::fmt::Formatter<'_>) -> std::fmt::Result {
        f.write_fmt(format_args!(
            "\tLaunch Template ID: {}\n",
            self.launch_template_arn
        ))?;
        f.write_fmt(format_args!(
            "\tScaling Group Name: {}\n",
            self.auto_scaling_group_name
        ))?;

        Ok(())
    }
}

pub struct AutoScalingScenarioDescription {
    group: Result<Vec<String>, ScenarioError>,
    instances: Result<Vec<String>, anyhow::Error>,
    activities: Result<Vec<Activity>, anyhow::Error>,
}

impl Display for AutoScalingScenarioDescription {
    fn fmt(&self, f: &mut std::fmt::Formatter<'_>) -> std::fmt::Result {
        writeln!(f, "\t      Group status:")?;
        match &self.group {
            Ok(groups) => {
                for status in groups {
                    writeln!(f, "\t\t- {status}")?;
                }
            }
```

```
            Err(e) => writeln!(f, "\t\t! - {e}")?,
        }
        writeln!(f, "\t         Instances:")?;
        match &self.instances {
            Ok(instances) => {
                for instance in instances {
                    writeln!(f, "\t\t- {instance}")?;
                }
            }
            Err(e) => writeln!(f, "\t\t! {e}")?,
        }

        writeln!(f, "\t        Activities:")?;
        match &self.activities {
            Ok(activities) => {
                for activity in activities {
                    writeln!(
                        f,
                        "\t\t- {} Progress: {}% Status: {:?} End: {:?}",
                        activity.cause().unwrap_or("Unknown"),
                        activity.progress.unwrap_or(-1),
                        activity.status_code(),
                        // activity.status_message().unwrap_or_default()
                        activity.end_time(),
                    )?;
                }
            }
            Err(e) => writeln!(f, "\t\t! {e}")?,
        }

        Ok(())
    }
}

#[derive(Debug)]
struct MetadataError {
    message: Option<String>,
    code: Option<String>,
}

impl MetadataError {
    fn from(err: &dyn ProvideErrorMetadata) -> Self {
        MetadataError {
            message: err.message().map(|s| s.to_string()),
```

```
            code: err.code().map(|s| s.to_string()),
        }
    }
}

impl Display for MetadataError {
    fn fmt(&self, f: &mut std::fmt::Formatter<'_>) -> std::fmt::Result {
        let display = match (&self.message, &self.code) {
            (None, None) => "Unknown".to_string(),
            (None, Some(code)) => format!("({code})"),
            (Some(message), None) => message.to_string(),
            (Some(message), Some(code)) => format!("{message} ({code})"),
        };
        write!(f, "{display}")
    }
}

#[derive(Debug)]
pub struct ScenarioError {
    message: String,
    context: Option<MetadataError>,
}

impl ScenarioError {
    pub fn with(message: impl Into<String>) -> Self {
        ScenarioError {
            message: message.into(),
            context: None,
        }
    }

    pub fn new(message: impl Into<String>, err: &dyn ProvideErrorMetadata) ->
 Self {
        ScenarioError {
            message: message.into(),
            context: Some(MetadataError::from(err)),
        }
    }
}

impl Error for ScenarioError {
    // While `Error` can capture `source` information about the underlying error,
 for this example
```

```
    // the ScenarioError captures the underlying information in MetadataError and
 treats it as a
    // single Error from this Crate. In other contexts, it may be appropriate to
 model the error
    // as including the SdkError as its source.
}
impl Display for ScenarioError {
    fn fmt(&self, f: &mut std::fmt::Formatter<'_>) -> std::fmt::Result {
        match &self.context {
            Some(c) => write!(f, "{}: {}", self.message, c),
            None => write!(f, "{}", self.message),
        }
    }
}

impl AutoScalingScenario {
    pub async fn prepare_scenario(sdk_config: &SdkConfig) -> Result<Self,
 Vec<ScenarioError>> {
        let ec2 = aws_sdk_ec2::Client::new(sdk_config);
        let autoscaling = aws_sdk_autoscaling::Client::new(sdk_config);

        let auto_scaling_group_name = String::from(AUTOSCALING_GROUP_NAME);

        // Before creating any resources, prepare the list of AZs
        let availablity_zones = ec2.describe_availability_zones().send().await;
        if let Err(err) = availablity_zones {
            return Err(vec![ScenarioError::new("Failed to find AZs", &err)]);
        }

        let availability_zones: Vec<String> = availablity_zones
            .unwrap()
            .availability_zones
            .unwrap_or_default()
            .iter()
            .take(3)
            .map(|z| z.zone_name.clone().unwrap())
            .collect();

        // 1. Create an EC2 launch template that you'll use to create an auto
 scaling group. Bonus: use SDK with EC2.CreateLaunchTemplate to create the launch
 template.
        //   * Recommended: InstanceType='t1.micro',
 ImageId='ami-0ca285d4c2cda3300'
        let create_launch_template = ec2
```

```
            .create_launch_template()
            .launch_template_name(LAUNCH_TEMPLATE_NAME)
            .launch_template_data(
                RequestLaunchTemplateData::builder()
                    .instance_type(aws_sdk_ec2::types::InstanceType::T1Micro)
                    .image_id("ami-0ca285d4c2cda3300")
                    .build(),
            )
            .send()
            .await
            .map_err(|err| vec![ScenarioError::new("Failed to create launch
template", &err)])?;

        let launch_template_arn = match create_launch_template.launch_template {
            Some(launch_template) =>
launch_template.launch_template_id.unwrap_or_default(),
            None => {
                // Try to delete the launch template
                let _ = ec2
                    .delete_launch_template()
                    .launch_template_name(LAUNCH_TEMPLATE_NAME)
                    .send()
                    .await;
                return Err(vec![ScenarioError::with("Failed to load launch
template")]);
            }
        };

        // 2. CreateAutoScalingGroup: pass it the launch template you created in
step 0. Give it min/max of 1 instance.
        //   You can use EC2.describe_availability_zones() to get a list of AZs
(you have to specify an AZ when you create the group).
        //   Wait for instance to launch. Use a waiter if you have one, otherwise
DescribeAutoScalingInstances until LifecycleState='InService'
        if let Err(err) = autoscaling
            .create_auto_scaling_group()
            .auto_scaling_group_name(auto_scaling_group_name.as_str())
            .launch_template(
                LaunchTemplateSpecification::builder()
                    .launch_template_id(launch_template_arn.clone())
                    .version("$Latest")
                    .build(),
            )
            .max_size(1)
```

```
            .min_size(1)
            .set_availability_zones(Some(availability_zones))
            .send()
            .await
        {
            let mut errs = vec![ScenarioError::new(
                "Failed to create autoscaling group",
                &err,
            )];

            if let Err(err) = autoscaling
                .delete_auto_scaling_group()
                .auto_scaling_group_name(auto_scaling_group_name.as_str())
                .send()
                .await
            {
                errs.push(ScenarioError::new(
                    "Failed to clean up autoscaling group",
                    &err,
                ));
            }

            if let Err(err) = ec2
                .delete_launch_template()
                .launch_template_id(launch_template_arn.clone())
                .send()
                .await
            {
                errs.push(ScenarioError::new(
                    "Failed to clean up launch template",
                    &err,
                ));
            }
            return Err(errs);
        }

        let scenario = AutoScalingScenario {
            ec2,
            autoscaling: autoscaling.clone(), // Clients are cheap so cloning
here to prevent a move is ok.
            auto_scaling_group_name: auto_scaling_group_name.clone(),
            launch_template_arn,
        };
```

```
        let enable_metrics_collection = autoscaling
            .enable_metrics_collection()
            .auto_scaling_group_name(auto_scaling_group_name.as_str())
            .granularity("1Minute")
            .set_metrics(Some(vec![
                String::from("GroupMinSize"),
                String::from("GroupMaxSize"),
                String::from("GroupDesiredCapacity"),
                String::from("GroupInServiceInstances"),
                String::from("GroupTotalInstances"),
            ]))
            .send()
            .await;

        match enable_metrics_collection {
            Ok(_) => Ok(scenario),
            Err(err) => {
                scenario.clean_scenario().await?;
                Err(vec![ScenarioError::new(
                    "Failed to enable metrics collections for group",
                    &err,
                )])
            }
        }
    }

    pub async fn clean_scenario(self) -> Result<(), Vec<ScenarioError>> {
        let _ = self.wait_for_no_scaling().await;
        let delete_group = self
            .autoscaling
            .delete_auto_scaling_group()
            .auto_scaling_group_name(self.auto_scaling_group_name.clone())
            .send()
            .await;

        // 14. Delete LaunchTemplate.
        let delete_launch_template = self
            .ec2
            .delete_launch_template()
            .launch_template_id(self.launch_template_arn.clone())
            .send()
            .await;

        let early_exit = match (delete_group, delete_launch_template) {
```

```
            (Ok(_), Ok(_)) => Ok(()),
            (Ok(_), Err(e)) => Err(vec![ScenarioError::new(
                "There was an error cleaning the launch template",
                &e,
            )]),
            (Err(e), Ok(_)) => Err(vec![ScenarioError::new(
                "There was an error cleaning the scale group",
                &e,
            )]),
            (Err(e1), Err(e2)) => Err(vec![
                ScenarioError::new("Multiple error cleaning the scenario Scale
Group", &e1),
                ScenarioError::new("Multiple error cleaning the scenario Launch
Template", &e2),
            ]),
        };

        if early_exit.is_err() {
            early_exit
        } else {
            // Wait for delete_group to finish
            let waiter = Waiter::new();
            let mut errors = Vec::<ScenarioError>::new();
            while errors.len() < 3 {
                if let Err(e) = waiter.sleep().await {
                    errors.push(e);
                    continue;
                }
                let describe_group = self
                    .autoscaling
                    .describe_auto_scaling_groups()

.auto_scaling_group_names(self.auto_scaling_group_name.clone())
                    .send()
                    .await;
                match describe_group {
                    Ok(group) => match group.auto_scaling_groups().first() {
                        Some(group) => {
                            if group.status() != Some("Delete in progress") {
                                errors.push(ScenarioError::with(format!(
                                    "Group in an unknown state while deleting:
{}",
                                    group.status().unwrap_or("unknown error")
                                )));
```

```
                            return Err(errors);
                        }
                    }
                    None => return Ok(()),
                },
                Err(err) => {
                    errors.push(ScenarioError::new("Failed to describe
autoscaling group during cleanup 3 times, last error", &err));
                }
            }
            if errors.len() > 3 {
                return Err(errors);
            }
        }
        Err(vec![ScenarioError::with(
            "Exited cleanup wait loop without retuning success or failing
after three rounds",
        )])
    }
}

pub async fn describe_scenario(&self) -> AutoScalingScenarioDescription {
    let group = self
        .autoscaling
        .describe_auto_scaling_groups()
        .auto_scaling_group_names(self.auto_scaling_group_name.clone())
        .send()
        .await
        .map(|s| {
            s.auto_scaling_groups()
                .iter()
                .map(|s| {
                    format!(
                        "{}: {}",
                        s.auto_scaling_group_name().unwrap_or("Unknown"),
                        s.status().unwrap_or("Unknown")
                    )
                })
                .collect::<Vec<String>>()
        })
        .map_err(|e| {
            ScenarioError::new("Failed to describe auto scaling groups for
scenario", &e)
        });
```

```
        let instances = self
            .list_instances()
            .await
            .map_err(|e| anyhow!("There was an error listing instances: {e}",));

        // 10. DescribeScalingActivities: list the scaling activities that have
occurred for the group so far.
        //   Bonus: use CloudWatch API to get and show some metrics collected for
the group.
        //   CW.ListMetrics with Namespace='AWS/AutoScaling' and
Dimensions=[{'Name': 'AutoScalingGroupName', 'Value': }]
        //   CW.GetMetricStatistics with Statistics='Sum'. Start and End times
must be in UTC!
        let activities = self
            .autoscaling
            .describe_scaling_activities()
            .auto_scaling_group_name(self.auto_scaling_group_name.clone())
            .into_paginator()
            .items()
            .send()
            .collect::<Result<Vec<_>, _>>()
            .await
            .map_err(|e| {
                anyhow!(
                    "There was an error retrieving scaling activities: {}",
                    DisplayErrorContext(&e)
                )
            });

        AutoScalingScenarioDescription {
            group,
            instances,
            activities,
        }
    }

    async fn get_group(&self) -> Result<AutoScalingGroup, ScenarioError> {
        let describe_auto_scaling_groups = self
            .autoscaling
            .describe_auto_scaling_groups()
            .auto_scaling_group_names(self.auto_scaling_group_name.clone())
            .send()
            .await;
```

```
        if let Err(err) = describe_auto_scaling_groups {
            return Err(ScenarioError::new(
                format!(
                    "Failed to get status of autoscaling group {}",
                    self.auto_scaling_group_name.clone()
                )
                .as_str(),
                &err,
            ));
        }

        let describe_auto_scaling_groups_output =
describe_auto_scaling_groups.unwrap();
        let auto_scaling_groups =
describe_auto_scaling_groups_output.auto_scaling_groups();
        let auto_scaling_group = auto_scaling_groups.first();

        if auto_scaling_group.is_none() {
            return Err(ScenarioError::with(format!(
                "Could not find autoscaling group {}",
                self.auto_scaling_group_name.clone()
            )));
        }

        Ok(auto_scaling_group.unwrap().clone())
    }

    pub async fn wait_for_no_scaling(&self) -> Result<(), ScenarioError> {
        let waiter = Waiter::new();
        let mut scaling = true;
        while scaling {
            waiter.sleep().await?;
            let describe_activities = self
                .autoscaling
                .describe_scaling_activities()
                .auto_scaling_group_name(self.auto_scaling_group_name.clone())
                .send()
                .await
                .map_err(|e| {
                    ScenarioError::new("Failed to get autoscaling activities for
group", &e)
                })?;
            let activities = describe_activities.activities();
```

```
            trace!(
                "Waiting for no scaling found {} activities",
                activities.len()
            );
            scaling = activities.iter().any(|a| a.progress() < Some(100));
        }
        Ok(())
    }

    pub async fn wait_for_stable(&self, size: usize) -> Result<(), ScenarioError>
 {
        self.wait_for_no_scaling().await?;

        let mut group = self.get_group().await?;
        let mut count = count_group_instances(&group);

        let waiter = Waiter::new();
        while count != size {
            trace!("Waiting for stable {size} (current: {count})");
            waiter.sleep().await?;
            group = self.get_group().await?;
            count = count_group_instances(&group);
        }

        Ok(())
    }

    pub async fn list_instances(&self) -> Result<Vec<String>, ScenarioError> {
        // The direct way to list instances is by using
 DescribeAutoScalingGroup's instances property. However, this returns a
 Vec<Instance>, as opposed to a Vec<AutoScalingInstanceDetails>.
        // Ok(self.get_group().await?.instances.unwrap_or_default().map(|
i| i.instance_id.clone().unwrap_or_default()).filter(|id| !
id.is_empty()).collect())

        // Alternatively, and for the sake of example,
 DescribeAutoScalingInstances returns a list that can be filtered by the client.
        self.autoscaling
            .describe_auto_scaling_instances()
            .into_paginator()
            .items()
            .send()
            .try_collect()
            .await
```

```
            .map(|items| {
                items
                    .into_iter()
                    .filter(|i| {
                        i.auto_scaling_group_name.as_deref()
                            == Some(self.auto_scaling_group_name.as_str())
                    })
                    .map(|i| i.instance_id.unwrap_or_default())
                    .filter(|id| !id.is_empty())
                    .collect::<Vec<String>>()
            })
            .map_err(|err| ScenarioError::new("Failed to get list of auto scaling
instances", &err))
    }

    pub async fn scale_min_size(&self, size: i32) -> Result<(), ScenarioError> {
        let update_group = self
            .autoscaling
            .update_auto_scaling_group()
            .auto_scaling_group_name(self.auto_scaling_group_name.clone())
            .min_size(size)
            .send()
            .await;
        if let Err(err) = update_group {
            return Err(ScenarioError::new(
                format!("Failer to update group to min size ({size}))").as_str(),
                &err,
            ));
        }
        Ok(())
    }

    pub async fn scale_max_size(&self, size: i32) -> Result<(), ScenarioError> {
        // 5. UpdateAutoScalingGroup: update max size to 3.
        let update_group = self
            .autoscaling
            .update_auto_scaling_group()
            .auto_scaling_group_name(self.auto_scaling_group_name.clone())
            .max_size(size)
            .send()
            .await;
        if let Err(err) = update_group {
            return Err(ScenarioError::new(
                format!("Failed to update group to max size ({size})").as_str(),
```

```
                &err,
            ));
        }
        Ok(())
    }

    pub async fn scale_desired_capacity(&self, capacity: i32) -> Result<(),
ScenarioError> {
        // 7. SetDesiredCapacity: set desired capacity to 2.
        //   Wait for a second instance to launch.
        let update_group = self
            .autoscaling
            .set_desired_capacity()
            .auto_scaling_group_name(self.auto_scaling_group_name.clone())
            .desired_capacity(capacity)
            .send()
            .await;
        if let Err(err) = update_group {
            return Err(ScenarioError::new(
                format!("Failed to update group to desired capacity
({capacity}))").as_str(),
                &err,
            ));
        }
        Ok(())
    }

    pub async fn scale_group_to_zero(&self) -> Result<(), ScenarioError> {
        // If this fails it's fine, just means there are extra cloudwatch metrics
events for the scale-down.
        let _ = self
            .autoscaling
            .disable_metrics_collection()
            .auto_scaling_group_name(self.auto_scaling_group_name.clone())
            .send()
            .await;

        // 12. DeleteAutoScalingGroup (to delete the group you must stop all
instances):
        //   UpdateAutoScalingGroup with MinSize=0
        let update_group = self
            .autoscaling
            .update_auto_scaling_group()
            .auto_scaling_group_name(self.auto_scaling_group_name.clone())
```

```
            .min_size(0)
            .desired_capacity(0)
            .send()
            .await;
        if let Err(err) = update_group {
            return Err(ScenarioError::new(
                "Failed to update group for scaling down&",
                &err,
            ));
        }

        let stable = self.wait_for_stable(0).await;
        if let Err(err) = stable {
            return Err(ScenarioError::with(format!(
                "Error while waiting for group to be stable on scale down: {err}"
            )));
        }

        Ok(())
    }

    pub async fn terminate_some_instance(&self) -> Result<(), ScenarioError> {
        // Retrieve a list of instances in the auto scaling group.
        let auto_scaling_group = self.get_group().await?;
        let instances = auto_scaling_group.instances();
        // Or use other logic to find an instance to terminate.
        let instance = instances.first();
        if let Some(instance) = instance {
            let instance_id = if let Some(instance_id) = instance.instance_id() {
                instance_id
            } else {
                return Err(ScenarioError::with("Missing instance id"));
            };
            let termination = self
                .ec2
                .terminate_instances()
                .instance_ids(instance_id)
                .send()
                .await;
            if let Err(err) = termination {
                Err(ScenarioError::new(
                    "There was a problem terminating an instance",
                    &err,
                ))
```

```
            } else {
                Ok(())
            }
        } else {
            Err(ScenarioError::with("There was no instance to terminate"))
        }
    }
}

fn count_group_instances(group: &AutoScalingGroup) -> usize {
    group.instances.as_ref().map(|i| i.len()).unwrap_or(0)
}
```

- Para obter detalhes da API, consulte os tópicos a seguir na Referência da API AWS SDK para Rust.
  - CreateAutoScalingGroup
  - DeleteAutoScalingGroup
  - DescribeAutoScalingGroups
  - DescribeAutoScalingInstances
  - DescribeScalingActivities
  - DisableMetricsCollection
  - EnableMetricsCollection
  - SetDesiredCapacity
  - TerminateInstanceInAutoScalingGroup
  - UpdateAutoScalingGroup

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Ações para Auto Scaling usando AWS SDKs

Os exemplos de código a seguir demonstram como realizar ações individuais do Auto Scaling com. AWS SDKs Cada exemplo inclui um link para GitHub, onde você pode encontrar instruções para configurar e executar o código.

Esses trechos chamam a API do Auto Scaling e são trechos de código de programas maiores que devem ser executados no contexto. É possível ver as ações em contexto em Cenários para o Auto Scaling usando AWS SDKs.

Os exemplos a seguir incluem apenas as ações mais utilizadas. Para obter uma lista completa, consulte a Referência da API Amazon EC2 Auto Scaling.

Exemplos

- Usar AttachInstances com uma CLI
- Use AttachLoadBalancerTargetGroups com um AWS SDK ou CLI
- Usar AttachLoadBalancers com uma CLI
- Usar CompleteLifecycleAction com uma CLI
- Use CreateAutoScalingGroup com um AWS SDK ou CLI
- Usar CreateLaunchConfiguration com uma CLI
- Usar CreateOrUpdateTags com uma CLI
- Use DeleteAutoScalingGroup com um AWS SDK ou CLI
- Usar DeleteLaunchConfiguration com uma CLI
- Usar DeleteLifecycleHook com uma CLI
- Usar DeleteNotificationConfiguration com uma CLI
- Usar DeletePolicy com uma CLI
- Usar DeleteScheduledAction com uma CLI
- Usar DeleteTags com uma CLI
- Usar DescribeAccountLimits com uma CLI
- Usar DescribeAdjustmentTypes com uma CLI
- Use DescribeAutoScalingGroups com um AWS SDK ou CLI
- Use DescribeAutoScalingInstances com um AWS SDK ou CLI
- Usar DescribeAutoScalingNotificationTypes com uma CLI
- Usar DescribeLaunchConfigurations com uma CLI
- Usar DescribeLifecycleHookTypes com uma CLI
- Usar DescribeLifecycleHooks com uma CLI
- Usar DescribeLoadBalancers com uma CLI
- Usar DescribeMetricCollectionTypes com uma CLI

- Usar DescribeNotificationConfigurations com uma CLI
- Usar DescribePolicies com uma CLI
- Use DescribeScalingActivities com um AWS SDK ou CLI
- Usar DescribeScalingProcessTypes com uma CLI
- Usar DescribeScheduledActions com uma CLI
- Usar DescribeTags com uma CLI
- Usar DescribeTerminationPolicyTypes com uma CLI
- Usar DetachInstances com uma CLI
- Usar DetachLoadBalancers com uma CLI
- Use DisableMetricsCollection com um AWS SDK ou CLI
- Use EnableMetricsCollection com um AWS SDK ou CLI
- Usar EnterStandby com uma CLI
- Usar ExecutePolicy com uma CLI
- Usar ExitStandby com uma CLI
- Usar PutLifecycleHook com uma CLI
- Usar PutNotificationConfiguration com uma CLI
- Usar PutScalingPolicy com uma CLI
- Usar PutScheduledUpdateGroupAction com uma CLI
- Usar RecordLifecycleActionHeartbeat com uma CLI
- Usar ResumeProcesses com uma CLI
- Use SetDesiredCapacity com um AWS SDK ou CLI
- Usar SetInstanceHealth com uma CLI
- Usar SetInstanceProtection com uma CLI
- Usar SuspendProcesses com uma CLI
- Use TerminateInstanceInAutoScalingGroup com um AWS SDK ou CLI
- Use UpdateAutoScalingGroup com um AWS SDK ou CLI

Usar **`AttachInstances`** com uma CLI

Os exemplos de código a seguir mostram como usar o `AttachInstances`.

CLI

AWS CLI

Para anexar uma instância a um grupo do Auto Scaling

Este exemplo anexa a instância especificada ao grupo do Auto Scaling especificado.

```
aws autoscaling attach-instances \
    --instance-ids i-061c63c5eb45f0416 \
    --auto-scaling-group-name my-asg
```

Este comando não produz saída.

- Para obter detalhes da API, consulte AttachInstancesem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo anexa o grupo de destino especificado ao grupo do Auto Scaling especificado. O Auto Scaling aumenta automaticamente a capacidade desejada do grupo do Auto Scaling.

```
Mount-ASInstance -InstanceId i-93633f9b -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte AttachInstancesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo anexa o grupo de destino especificado ao grupo do Auto Scaling especificado. O Auto Scaling aumenta automaticamente a capacidade desejada do grupo do Auto Scaling.

```
Mount-ASInstance -InstanceId i-93633f9b -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte AttachInstancesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Use **`AttachLoadBalancerTargetGroups`** com um AWS SDK ou CLI

Os exemplos de código a seguir mostram como usar o `AttachLoadBalancerTargetGroups`.

Exemplos de ações são trechos de código de programas maiores e devem ser executados em contexto. É possível ver essa ação em contexto no seguinte exemplo de código:

- Criar e gerenciar um serviço resiliente

.NET

SDK para .NET

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    /// <summary>
    /// Attaches an Elastic Load Balancing (ELB) target group to this EC2 Auto
 Scaling group.
    /// The
    /// </summary>
    /// <param name="autoScalingGroupName">The name of the Auto Scaling group.</
param>
    /// <param name="targetGroupArn">The Arn for the target group.</param>
    /// <returns>Async task.</returns>
    public async Task AttachLoadBalancerToGroup(string autoScalingGroupName,
 string targetGroupArn)
    {
        await _amazonAutoScaling.AttachLoadBalancerTargetGroupsAsync(
            new AttachLoadBalancerTargetGroupsRequest()
            {
                AutoScalingGroupName = autoScalingGroupName,
                TargetGroupARNs = new List<string>() { targetGroupArn }
            });
```

```
    }
```

- Para obter detalhes da API, consulte AttachLoadBalancerTargetGroupsa Referência AWS SDK para .NET da API.

CLI

AWS CLI

Como anexar um grupo de destino a um grupo do Auto Scaling

Este exemplo anexa o grupo de destino especificado ao grupo do Auto Scaling especificado.

```
aws autoscaling attach-load-balancer-target-groups \
    --auto-scaling-group-name my-asg \
    --target-group-arns arn:aws:elasticloadbalancing:us-
west-2:123456789012:targetgroup/my-targets/73e2d6bc24d8a067
```

Este comando não produz saída.

Para obter mais informações, consulte Elastic Load Balancing e Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte AttachLoadBalancerTargetGroupsem Referência de AWS CLI Comandos.

JavaScript

SDK para JavaScript (v3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    const client = new AutoScalingClient({});
    await client.send(
      new AttachLoadBalancerTargetGroupsCommand({
        AutoScalingGroupName: NAMES.autoScalingGroupName,
```

```
        TargetGroupARNs: [state.targetGroupArn],
      }),
    );
```

- Para obter detalhes da API, consulte AttachLoadBalancerTargetGroupsa Referência AWS SDK para JavaScript da API.

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """
    Encapsulates Amazon EC2 Auto Scaling and EC2 management actions.
    """

    def __init__(
        self,
        resource_prefix: str,
        inst_type: str,
        ami_param: str,
        autoscaling_client: boto3.client,
        ec2_client: boto3.client,
        ssm_client: boto3.client,
        iam_client: boto3.client,
    ):
        """
        Initializes the AutoScaler class with the necessary parameters.

        :param resource_prefix: The prefix for naming AWS resources that are
 created by this class.
        :param inst_type: The type of EC2 instance to create, such as t3.micro.
        :param ami_param: The Systems Manager parameter used to look up the AMI
 that is created.
```

```
        :param autoscaling_client: A Boto3 EC2 Auto Scaling client.
        :param ec2_client: A Boto3 EC2 client.
        :param ssm_client: A Boto3 Systems Manager client.
        :param iam_client: A Boto3 IAM client.
        """
        self.inst_type = inst_type
        self.ami_param = ami_param
        self.autoscaling_client = autoscaling_client
        self.ec2_client = ec2_client
        self.ssm_client = ssm_client
        self.iam_client = iam_client
        sts_client = boto3.client("sts")
        self.account_id = sts_client.get_caller_identity()["Account"]

        self.key_pair_name = f"{resource_prefix}-key-pair"
        self.launch_template_name = f"{resource_prefix}-template-"
        self.group_name = f"{resource_prefix}-group"

        # Happy path
        self.instance_policy_name = f"{resource_prefix}-pol"
        self.instance_role_name = f"{resource_prefix}-role"
        self.instance_profile_name = f"{resource_prefix}-prof"

        # Failure mode
        self.bad_creds_policy_name = f"{resource_prefix}-bc-pol"
        self.bad_creds_role_name = f"{resource_prefix}-bc-role"
        self.bad_creds_profile_name = f"{resource_prefix}-bc-prof"


    def attach_load_balancer_target_group(
        self, lb_target_group: Dict[str, Any]
    ) -> None:
        """
        Attaches an Elastic Load Balancing (ELB) target group to this EC2 Auto
Scaling group.
        The target group specifies how the load balancer forwards requests to the
instances
        in the group.

        :param lb_target_group: Data about the ELB target group to attach.
        """
        try:
            self.autoscaling_client.attach_load_balancer_target_groups(
                AutoScalingGroupName=self.group_name,
```

```
                TargetGroupARNs=[lb_target_group["TargetGroupArn"]],
            )
            log.info(
                "Attached load balancer target group %s to auto scaling group
%s.",
                lb_target_group["TargetGroupName"],
                self.group_name,
            )
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(
                f"Failed to attach load balancer target group
'{lb_target_group['TargetGroupName']}'."
            )
            if error_code == "ResourceContentionFault":
                log.error(
                    "The request failed due to a resource contention issue. "
                    "Ensure that no conflicting operations are being performed on
the resource."
                )
            elif error_code == "ServiceLinkedRoleFailure":
                log.error(
                    "The operation failed because the service-linked role is not
ready or does not exist. "
                    "Check that the service-linked role exists and is correctly
configured."
                )
            log.error(f"Full error:\n\t{err}")
```

- Para obter detalhes da API, consulte a AttachLoadBalancerTargetGroupsReferência da API AWS SDK for Python (Boto3).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`AttachLoadBalancers`** com uma CLI

Os exemplos de código a seguir mostram como usar o `AttachLoadBalancers`.

CLI

AWS CLI

Para anexar um Classic Load Balancer ao grupo do Auto Scaling

Este exemplo anexa o Classic Load Balancer especificado ao grupo do Auto Scaling especificado.

```
aws autoscaling attach-load-balancers \
    --load-balancer-names my-load-balancer \
    --auto-scaling-group-name my-asg
```

Este comando não produz saída.

Para obter mais informações, consulte Elastic Load Balancing e Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte AttachLoadBalancersem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo anexa o Balanceador de Carga especificado ao grupo do Auto Scaling especificado.

```
Add-ASLoadBalancer -LoadBalancerName my-lb -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte AttachLoadBalancersem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo anexa o Balanceador de Carga especificado ao grupo do Auto Scaling especificado.

```
Add-ASLoadBalancer -LoadBalancerName my-lb -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte AttachLoadBalancersem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`CompleteLifecycleAction`** com uma CLI

Os exemplos de código a seguir mostram como usar o `CompleteLifecycleAction`.

CLI

AWS CLI

Para concluir uma ação do ciclo de vida

Este exemplo notifica o Amazon EC2 Auto Scaling de que a ação do ciclo de vida especificada foi concluída para que ele possa concluir a execução ou o encerramento da instância.

```
aws autoscaling complete-lifecycle-action \
    --lifecycle-hook-name my-launch-hook \
    --auto-scaling-group-name my-asg \
    --lifecycle-action-result CONTINUE \
    --lifecycle-action-token bcd2f1b8-9a78-44d3-8a7a-4dd07d7cf635
```

Este comando não produz saída.

Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte CompleteLifecycleActionem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo conclui a ação de ciclo de vida especificada.

```
Complete-ASLifecycleAction -LifecycleHookName myLifecycleHook -
AutoScalingGroupName my-asg -LifecycleActionResult CONTINUE -LifecycleActionToken
 bcd2f1b8-9a78-44d3-8a7a-4dd07d7cf635
```

- Para obter detalhes da API, consulte CompleteLifecycleActionem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo conclui a ação de ciclo de vida especificada.

```
Complete-ASLifecycleAction -LifecycleHookName myLifecycleHook -
AutoScalingGroupName my-asg -LifecycleActionResult CONTINUE -LifecycleActionToken
 bcd2f1b8-9a78-44d3-8a7a-4dd07d7cf635
```

- Para obter detalhes da API, consulte CompleteLifecycleActionem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Use **`CreateAutoScalingGroup`** com um AWS SDK ou CLI

Os exemplos de código a seguir mostram como usar o `CreateAutoScalingGroup`.

Exemplos de ações são trechos de código de programas maiores e devem ser executados em contexto. É possível ver essa ação em contexto nos seguintes exemplos de código:

- Conheça os conceitos básicos
- Criar e gerenciar um serviço resiliente

.NET

SDK para .NET (v4)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    /// <summary>
```

```
    /// Create a new Amazon EC2 Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name to use for the new Auto Scaling
    /// group.</param>
    /// <param name="launchTemplateName">The name of the Amazon EC2 Auto Scaling
    /// launch template to use to create instances in the group.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> CreateAutoScalingGroupAsync(
        string groupName,
        string launchTemplateName,
        string availabilityZone)
    {
        var templateSpecification = new LaunchTemplateSpecification
        {
            LaunchTemplateName = launchTemplateName,
        };

        var zoneList = new List<string>
            {
                availabilityZone,
            };

        var request = new CreateAutoScalingGroupRequest
        {
            AutoScalingGroupName = groupName,
            AvailabilityZones = zoneList,
            LaunchTemplate = templateSpecification,
            MaxSize = 6,
            MinSize = 1
        };
        try
        {
            var response = await
_amazonAutoScaling.CreateAutoScalingGroupAsync(request);
            Console.WriteLine($"{groupName} Auto Scaling Group created");
            return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
        }
        catch (AlreadyExistsException)
        {
            Console.WriteLine($"{groupName} Auto Scaling Group already exists.");
            return true;
        }
    }
```

- Para obter detalhes da API, consulte CreateAutoScalingGroupa Referência AWS SDK para .NET da API.

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

        Aws::AutoScaling::Model::CreateAutoScalingGroupRequest request;
        request.SetAutoScalingGroupName(groupName);
        Aws::Vector<Aws::String> availabilityGroupZones;
        availabilityGroupZones.push_back(
                availabilityZones[availabilityZoneChoice - 1].GetZoneName());
        request.SetAvailabilityZones(availabilityGroupZones);
        request.SetMaxSize(1);
        request.SetMinSize(1);

        Aws::AutoScaling::Model::LaunchTemplateSpecification
launchTemplateSpecification;
        launchTemplateSpecification.SetLaunchTemplateName(templateName);
        request.SetLaunchTemplate(launchTemplateSpecification);

        Aws::AutoScaling::Model::CreateAutoScalingGroupOutcome outcome =
                autoScalingClient.CreateAutoScalingGroup(request);

        if (outcome.IsSuccess()) {
            std::cout << "Created Auto Scaling group '" << groupName << "'..."
                      << std::endl;
```

```
        }
        else if (outcome.GetError().GetErrorType() ==
                 Aws::AutoScaling::AutoScalingErrors::ALREADY_EXISTS_FAULT) {
            std::cout << "Auto Scaling group '" << groupName << "' already
exists."
                      << std::endl;
        }
        else {
            std::cerr << "Error with AutoScaling::CreateAutoScalingGroup. "
                      << outcome.GetError().GetMessage()
                      << std::endl;

        }
```

- Para obter detalhes da API, consulte CreateAutoScalingGroupa Referência AWS SDK para C++ da API.

CLI

AWS CLI

Exemplo 1: como criar um grupo do Auto Scaling

O exemplo de `create-auto-scaling-group` a seguir cria um grupo do Auto Scaling em sub-redes de várias zonas de disponibilidade de uma região. As instâncias são executadas com a versão padrão do modelo de execução especificado. Observe que os padrões são usados na maioria das outras configurações, como nas políticas de encerramento e na configuração de verificação de integridade.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12 \
    --min-size 1 \
    --max-size 5 \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Este comando não produz saída.

Para obter mais informações, consulte Grupos do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: como anexar o Application Load Balancer, o Network Load Balancer ou o Gateway Load Balancer

Este exemplo especifica o ARN de um grupo de destino para um balanceador de carga compatível com o tráfego esperado. O tipo de verificação de integridade especifica o ELB. Desta forma, quando o Elastic Load Balancing reportar uma instância como não íntegra, o grupo do Auto Scaling a substitui. O comando também define um período de carência de 600 segundos para a verificação de integridade. O período de carência ajuda a evitar o encerramento prematuro de instâncias recém-iniciadas.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12 \
    --target-group-arns arn:aws:elasticloadbalancing:us-
west-2:123456789012:targetgroup/my-targets/943f017f100becff \
    --health-check-type ELB \
    --health-check-grace-period 600 \
    --min-size 1 \
    --max-size 5 \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Este comando não produz saída.

Para obter mais informações, consulte Elastic Load Balancing e Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 3: como especificar um grupo de posicionamento e usar a versão mais recente do modelo de execução

Este exemplo executa instâncias em um grupo de posicionamento dentro de uma única zona de disponibilidade. Isso pode ser útil para grupos de baixa latência com workloads de HPC. Esse exemplo também especifica o tamanho mínimo e máximo e a capacidade desejada do grupo.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12,Version='$Latest' \
    --min-size 1 \
    --max-size 5 \
    --desired-capacity 3 \
    --placement-group my-placement-group \
```

```
    --vpc-zone-identifier "subnet-6194ea3b"
```

Este comando não produz saída.

Para obter mais informações, consulte Grupos de posicionamento no Guia do usuário para instâncias do Linux do Amazon EC2.

Exemplo 4: como especificar um grupo do Auto Scaling de instância única e usar uma versão específica para iniciar o modelo

Este exemplo cria um grupo do Auto Scaling com capacidade mínima e máxima definida como 1 para impor que uma apenas instância seja executada. O comando também especifica a v1 de um modelo de execução no qual o ID de um ENI existente é especificado. Ao usar um modelo de execução que especifica um ENI existente para eth0, é necessário especificar uma zona de disponibilidade para o grupo do Auto Scaling que corresponda à interface de rede, mas sem especificar um ID de sub-rede na solicitação.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg-single-instance \
    --launch-template LaunchTemplateName=my-template-for-auto-scaling,Version='1'
 \
    --min-size 1 \
    --max-size 1 \
    --availability-zones us-west-2a
```

Este comando não produz saída.

Para obter mais informações, consulte Grupos do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 5: como especificar uma política de encerramento diferente

Este exemplo cria um grupo do Auto Scaling usando uma configuração de execução e define a política de encerramento para encerrar as instâncias mais antigas primeiro. O comando também aplica uma tag ao grupo e suas instâncias, com uma chave `Role` e valor de `WebServer`.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-configuration-name my-lc \
    --min-size 1 \
    --max-size 5 \
```

```
    --termination-policies "OldestInstance" \
    --tags "ResourceId=my-asg,ResourceType=auto-scaling-
group,Key=Role,Value=WebServer,PropagateAtLaunch=true" \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Este comando não produz saída.

Para obter mais informações, consulte Trabalhar com políticas de término do Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 6: como especificar um gancho do ciclo de vida de lançamento

Este exemplo a seguir cria um grupo do Auto Scaling com um gancho do ciclo de vida que oferece suporte a uma ação personalizada na inicialização da instância.

```
aws autoscaling create-auto-scaling-group \
    --cli-input-json file://~/config.json
```

Conteúdo do arquivo `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-1234567890abcde12"
    },
    "LifecycleHookSpecificationList": [{
        "LifecycleHookName": "my-launch-hook",
        "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
        "NotificationTargetARN": "arn:aws:sqs:us-west-2:123456789012:my-sqs-
queue",
        "RoleARN": "arn:aws:iam::123456789012:role/my-notification-role",
        "NotificationMetadata": "SQS message metadata",
        "HeartbeatTimeout": 4800,
        "DefaultResult": "ABANDON"
    }],
    "MinSize": 1,
    "MaxSize": 5,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
    "Tags": [{
        "ResourceType": "auto-scaling-group",
        "ResourceId": "my-asg",
        "PropagateAtLaunch": true,
        "Value": "test",
```

```
        "Key": "environment"
    }]
}
```

Este comando não produz saída.

Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 7: como especificar um gancho do ciclo de vida de encerramento

Este exemplo a seguir cria um grupo do Auto Scaling com um gancho do ciclo de vida que oferece suporte a uma ação personalizada no encerramento da instância.

```
aws autoscaling create-auto-scaling-group \
    --cli-input-json file://~/config.json
```

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-1234567890abcde12"
    },
    "LifecycleHookSpecificationList": [{
        "LifecycleHookName": "my-termination-hook",
        "LifecycleTransition": "autoscaling:EC2_INSTANCE_TERMINATING",
        "HeartbeatTimeout": 120,
        "DefaultResult": "CONTINUE"
    }],
    "MinSize": 1,
    "MaxSize": 5,
    "TargetGroupARNs": [
        "arn:aws:elasticloadbalancing:us-west-2:123456789012:targetgroup/my-
targets/73e2d6bc24d8a067"
    ],
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

Este comando não produz saída.

Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 8: como especificar uma política de encerramento personalizada

Este exemplo cria um grupo do Auto Scaling que especifica uma política de encerramento da função do Lambda personalizada que diz ao Amazon EC2 Auto Scaling quais instâncias podem ser encerradas com segurança em escala.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg-single-instance \
    --launch-template LaunchTemplateName=my-template-for-auto-scaling \
    --min-size 1 \
    --max-size 5 \
    --termination-policies "arn:aws:lambda:us-
west-2:123456789012:function:HelloFunction:prod" \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Este comando não produz saída.

Para obter mais informações, consulte Criar uma política de término personalizada com o Lambda no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte CreateAutoScalingGroupem Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
import software.amazon.awssdk.core.waiters.WaiterResponse;
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.autoscaling.AutoScalingClient;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingException;
import
 software.amazon.awssdk.services.autoscaling.model.CreateAutoScalingGroupRequest;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsRequest;
```

```
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsResponse;
import
 software.amazon.awssdk.services.autoscaling.model.LaunchTemplateSpecification;
import software.amazon.awssdk.services.autoscaling.waiters.AutoScalingWaiter;

/**
 * Before running this SDK for Java (v2) code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class CreateAutoScalingGroup {
    public static void main(String[] args) {
        final String usage = """

                Usage:
                    <groupName> <launchTemplateName> <serviceLinkedRoleARN>
 <vpcZoneId>

                Where:
                    groupName - The name of the Auto Scaling group.
                    launchTemplateName - The name of the launch template.\s
                    vpcZoneId - A subnet Id for a virtual private cloud (VPC)
 where instances in the Auto Scaling group can be created.
                """;

        if (args.length != 3) {
            System.out.println(usage);
            System.exit(1);
        }

        String groupName = args[0];
        String launchTemplateName = args[1];
        String vpcZoneId = args[2];
        AutoScalingClient autoScalingClient = AutoScalingClient.builder()
                .region(Region.US_EAST_1)
                .build();

        createAutoScalingGroup(autoScalingClient, groupName, launchTemplateName,
 vpcZoneId);
```

```
        autoScalingClient.close();
    }

    public static void createAutoScalingGroup(AutoScalingClient
 autoScalingClient,
            String groupName,
            String launchTemplateName,
            String vpcZoneId) {

        try {
            AutoScalingWaiter waiter = autoScalingClient.waiter();
            LaunchTemplateSpecification templateSpecification =
 LaunchTemplateSpecification.builder()
                    .launchTemplateName(launchTemplateName)
                    .build();

            CreateAutoScalingGroupRequest request =
 CreateAutoScalingGroupRequest.builder()
                    .autoScalingGroupName(groupName)
                    .availabilityZones("us-east-1a")
                    .launchTemplate(templateSpecification)
                    .maxSize(1)
                    .minSize(1)
                    .vpcZoneIdentifier(vpcZoneId)
                    .build();

            autoScalingClient.createAutoScalingGroup(request);
            DescribeAutoScalingGroupsRequest groupsRequest =
 DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();

            WaiterResponse<DescribeAutoScalingGroupsResponse> waiterResponse =
 waiter
                    .waitUntilGroupExists(groupsRequest);
            waiterResponse.matched().response().ifPresent(System.out::println);
            System.out.println("Auto Scaling Group created");

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- Para obter detalhes da API, consulte CreateAutoScalingGroupa Referência AWS SDK for Java 2.x da API.

Kotlin

SDK para Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun createAutoScalingGroup(
    groupName: String,
    launchTemplateNameVal: String,
    serviceLinkedRoleARNVal: String,
    vpcZoneIdVal: String,
) {
    val templateSpecification =
        LaunchTemplateSpecification {
            launchTemplateName = launchTemplateNameVal
        }

    val request =
        CreateAutoScalingGroupRequest {
            autoScalingGroupName = groupName
            availabilityZones = listOf("us-east-1a")
            launchTemplate = templateSpecification
            maxSize = 1
            minSize = 1
            vpcZoneIdentifier = vpcZoneIdVal
            serviceLinkedRoleArn = serviceLinkedRoleARNVal
        }

    // This object is required for the waiter call.
    val groupsRequestWaiter =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
```

```
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.createAutoScalingGroup(request)
        autoScalingClient.waitUntilGroupExists(groupsRequestWaiter)
        println("$groupName was created!")
    }
}
```

- Para obter detalhes da API, consulte a CreateAutoScalingGroupreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function createAutoScalingGroup(
        $autoScalingGroupName,
        $availabilityZones,
        $minSize,
        $maxSize,
        $launchTemplateId
    ) {
        return $this->autoScalingClient->createAutoScalingGroup([
            'AutoScalingGroupName' => $autoScalingGroupName,
            'AvailabilityZones' => $availabilityZones,
            'MinSize' => $minSize,
            'MaxSize' => $maxSize,
            'LaunchTemplate' => [
                'LaunchTemplateId' => $launchTemplateId,
            ],
        ]);
    }
```

- Para obter detalhes da API, consulte CreateAutoScalingGroupa Referência AWS SDK para PHP da API.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo cria um grupo do Auto Scaling com o nome e os atributos especificados. A capacidade desejada padrão é o tamanho mínimo. Portanto, esse grupo do Auto Scaling inicia duas instâncias, uma em cada uma das duas zonas de disponibilidade especificadas.

```
New-ASAutoScalingGroup -AutoScalingGroupName my-asg -LaunchConfigurationName my-
lc -MinSize 2 -MaxSize 6 -AvailabilityZone @("us-west-2a", "us-west-2b")
```

- Para obter detalhes da API, consulte CreateAutoScalingGroupem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo cria um grupo do Auto Scaling com o nome e os atributos especificados. A capacidade desejada padrão é o tamanho mínimo. Portanto, esse grupo do Auto Scaling inicia duas instâncias, uma em cada uma das duas zonas de disponibilidade especificadas.

```
New-ASAutoScalingGroup -AutoScalingGroupName my-asg -LaunchConfigurationName my-
lc -MinSize 2 -MaxSize 6 -AvailabilityZone @("us-west-2a", "us-west-2b")
```

- Para obter detalhes da API, consulte CreateAutoScalingGroupem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def create_group(
        self,
        group_name: str,
        group_zones: List[str],
        launch_template_name: str,
        min_size: int,
        max_size: int,
    ) -> None:
        """
        Creates an Auto Scaling group.

        :param group_name: The name to give to the group.
        :param group_zones: The Availability Zones in which instances can be
 created.
        :param launch_template_name: The name of an existing Amazon EC2 launch
 template.
                                     The launch template specifies the
 configuration of
                                     instances that are created by auto scaling
 activities.
        :param min_size: The minimum number of active instances in the group.
        :param max_size: The maximum number of active instances in the group.
        :return: None
        :raises ClientError: If there is an error creating the Auto Scaling
 group.
        """
        try:
            self.autoscaling_client.create_auto_scaling_group(
                AutoScalingGroupName=group_name,
                AvailabilityZones=group_zones,
                LaunchTemplate={
                    "LaunchTemplateName": launch_template_name,
                    "Version": "$Default",
```

```
                },
                MinSize=min_size,
                MaxSize=max_size,
            )

            # Wait for the group to exist.
            waiter = self.autoscaling_client.get_waiter("group_exists")
            waiter.wait(AutoScalingGroupNames=[group_name])

            logger.info(f"Successfully created Auto Scaling group {group_name}.")

        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            logger.error(f"Failed to create Auto Scaling group {group_name}.")
            if error_code == "AlreadyExistsFault":
                logger.error(
                    f"An Auto Scaling group with the name '{group_name}' already
 exists. "
                    "Please use a different name or update the existing group.",
                )
            elif error_code == "LimitExceededFault":
                logger.error(
                    "The request failed because you have reached the limit "
                    "on the number of Auto Scaling groups or launch
 configurations. "
                    "Consider deleting unused resources or request a limit
 increase. "
                    "\nSee Auto Scaling Service Quota documentation here:"
                    "\n\thttps://docs.aws.amazon.com/autoscaling/ec2/userguide/
ec2-auto-scaling-quotas.html"
                )
            logger.error(f"Full error:\n\t{err}")
            raise
```

- Para obter detalhes da API, consulte a CreateAutoScalingGroupReferência da API AWS SDK for Python (Boto3).

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
async fn create_group(client: &Client, name: &str, id: &str) -> Result<(), Error>
 {
    client
        .create_auto_scaling_group()
        .auto_scaling_group_name(name)
        .instance_id(id)
        .min_size(1)
        .max_size(5)
        .send()
        .await?;

    println!("Created AutoScaling group");

    Ok(())
}
```

- Para obter detalhes da API, consulte a CreateAutoScalingGroupreferência da API AWS SDK for Rust.

SAP ABAP

SDK para SAP ABAP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    DATA lo_launch_template TYPE REF TO /aws1/cl_asclaunchtemplatespec.

    " Example: iv_group_name = 'my-auto-scaling-group'
    " Example: iv_launch_template_name = 'my-launch-template'
    " Example: iv_min_size = 1
    " Example: iv_max_size = 3
    " Example: iv_vpc_zone_identifier = 'subnet-12345,subnet-67890' (for VPC)

    TRY.
        " Create launch template specification
        lo_launch_template = NEW /aws1/cl_asclaunchtemplatespec(
          iv_launchtemplatename = iv_launch_template_name
          iv_version = '$Default' ).

        " Create the Auto Scaling group
        " Use VPCZoneIdentifier for VPC subnets, or AvailabilityZones for EC2-
Classic
        IF iv_vpc_zone_identifier IS NOT INITIAL.
          " VPC-based deployment - use subnet IDs
          ao_asc->createautoscalinggroup(
            iv_autoscalinggroupname = iv_group_name
            iv_vpczoneidentifier = iv_vpc_zone_identifier
            io_launchtemplate = lo_launch_template
            iv_minsize = iv_min_size
            iv_maxsize = iv_max_size ).
        ELSE.
          " EC2-Classic or default VPC - use availability zones
          ao_asc->createautoscalinggroup(
            iv_autoscalinggroupname = iv_group_name
            it_availabilityzones = it_group_zones
            io_launchtemplate = lo_launch_template
            iv_minsize = iv_min_size
            iv_maxsize = iv_max_size ).
        ENDIF.

        " Wait for the group to be created (simplified - in production use proper
 polling)
        WAIT UP TO 10 SECONDS.

        MESSAGE 'Auto Scaling group created successfully' TYPE 'I'.

      CATCH /aws1/cx_ascalreadyexistsfault INTO DATA(lo_already_exists).
        RAISE EXCEPTION lo_already_exists.
```

```
      CATCH /aws1/cx_asclimitexceededfault INTO DATA(lo_limit_exceeded).
        RAISE EXCEPTION lo_limit_exceeded.
      CATCH /aws1/cx_rt_generic INTO DATA(lo_generic_exception).
        RAISE EXCEPTION lo_generic_exception.
    ENDTRY.
```

- Para obter detalhes da API, consulte a CreateAutoScalingGroupreferência da API AWS SDK for SAP ABAP.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`CreateLaunchConfiguration`** com uma CLI

Os exemplos de código a seguir mostram como usar o `CreateLaunchConfiguration`.

CLI

AWS CLI

Exemplo 1: Saiba como criar uma configuração de execução.

Este exemplo cria uma configuração de inicialização simples.

```
aws autoscaling create-launch-configuration \
    --launch-configuration-name my-lc \
    --image-id ami-04d5cc9b88example \
    --instance-type m5.large
```

Este comando não produz saída.

Para obter mais informações, consulte Criar uma configuração de execução no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: Para criar uma configuração de inicialização com um grupo de segurança, um key pair e um script de inicialização

Este exemplo cria uma configuração de inicialização com um grupo de segurança, um key pair e um script de inicialização contido nos dados do usuário.

```
aws autoscaling create-launch-configuration \
    --launch-configuration-name my-lc \
    --image-id ami-04d5cc9b88example \
    --instance-type m5.large \
    --security-groups sg-eb2af88example \
    --key-name my-key-pair \
    --user-data file://myuserdata.txt
```

Este comando não produz saída.

Para obter mais informações, consulte Criar uma configuração de execução no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 3: criar uma configuração de inicialização com um perfil do IAM

Este exemplo cria uma configuração de execução com o nome do perfil de instância de um perfil do IAM.

```
aws autoscaling create-launch-configuration \
    --launch-configuration-name my-lc \
    --image-id ami-04d5cc9b88example \
    --instance-type m5.large \
    --iam-instance-profile my-autoscaling-role
```

Este comando não produz saída.

Para obter mais informações, consulte Função do IAM para aplicações que são executadas em instâncias do Amazon EC2 no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 4: habilitar o monitoramento detalhado em uma configuração de execução

Este exemplo cria uma configuração de lançamento com o monitoramento detalhado do EC2 ativado, que envia métricas do EC2 CloudWatch em períodos de 1 minuto.

```
aws autoscaling create-launch-configuration \
    --launch-configuration-name my-lc \
    --image-id ami-04d5cc9b88example \
    --instance-type m5.large \
    --instance-monitoring Enabled=true
```

Este comando não produz saída.

Para obter mais informações, consulte Configurando o monitoramento de instâncias do Auto Scaling na Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 5: criar uma configuração de execução para instâncias spot

Este exemplo cria uma configuração de execução que usa instâncias spot como a única opção de compra.

```
aws autoscaling create-launch-configuration \
    --launch-configuration-name my-lc \
    --image-id ami-04d5cc9b88example \
    --instance-type m5.large \
    --spot-price "0.50"
```

Este comando não produz saída.

Para obter mais informações, consulte Solicitar instâncias spot no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 6: criar uma configuração de execução usando uma instância do EC2

Este exemplo cria uma configuração de execução com base nos atributos de uma instância existente. Ele substitui a locação de localização e determina se um endereço IP público é definido, incluindo as opções `--placement-tenancy` e `--no-associate-public-ip-address`.

```
aws autoscaling create-launch-configuration \
    --launch-configuration-name my-lc-from-instance \
    --instance-id i-0123a456700123456 \
    --instance-type m5.large \
    --no-associate-public-ip-address \
    --placement-tenancy dedicated
```

Este comando não produz saída.

Para obter mais informações, consulte Criação de uma configuração de execução usando uma instância do EC2 no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 7: criar uma configuração de execução com um mapeamento de dispositivos de blocos para um volume do Amazon EBS

Este exemplo cria uma configuração de execução com um mapeamento de dispositivos de blocos para um volume gp3 do Amazon EBS com o nome do dispositivo `/dev/sdh` e um tamanho de volume de 20.

```
aws autoscaling create-launch-configuration \
    --launch-configuration-name my-lc \
    --image-id ami-04d5cc9b88example \
    --instance-type m5.large \
    --block-device-mappings '[{"DeviceName":"/dev/sdh","Ebs":
{"VolumeSize":20,"VolumeType":"gp3"}}]'
```

Este comando não produz saída.

Para obter mais informações, consulte EBS na Referência da API do Amazon EC2 Auto Scaling.

Para obter informações sobre a sintaxe para citar valores de parâmetros formatados em JSON, consulte Usando aspas com cadeias de caracteres na CLI AWS no Guia do usuário da interface de linha de comando.AWS

Exemplo 8: criar uma configuração de execução com um mapeamento de dispositivos de blocos para um volume de armazenamento de instância

Este exemplo cria uma configuração de execução com `ephemeral1` como volume de armazenamento de instância com o nome do dispositivo `/dev/sdc`.

```
aws autoscaling create-launch-configuration \
    --launch-configuration-name my-lc \
    --image-id ami-04d5cc9b88example \
    --instance-type m5.large \
    --block-device-mappings '[{"DeviceName":"/dev/
sdc","VirtualName":"ephemeral1"}]'
```

Este comando não produz saída.

Para obter mais informações, consulte a BlockDeviceMappingReferência da API Amazon EC2 Auto Scaling.

Para obter informações sobre a sintaxe para citar valores de parâmetros formatados em JSON, consulte Usando aspas com cadeias de caracteres na CLI AWS no Guia do usuário da interface de linha de comando.AWS

Exemplo 9: criar uma configuração de execução e impedir que um dispositivo de blocos se conecte no momento da execução

Este exemplo cria uma configuração de execução que suprime um dispositivo de blocos especificado pelo mapeamento de dispositivos de blocos da AMI (por exemplo, `/dev/sdf`).

```
aws autoscaling create-launch-configuration \
    --launch-configuration-name my-lc \
    --image-id ami-04d5cc9b88example \
    --instance-type m5.large \
    --block-device-mappings '[{"DeviceName":"/dev/sdf","NoDevice":""}]'
```

Este comando não produz saída.

Para obter mais informações, consulte a BlockDeviceMappingReferência da API Amazon EC2 Auto Scaling.

Para obter informações sobre a sintaxe para citar valores de parâmetros formatados em JSON, consulte Usando aspas com cadeias de caracteres na CLI AWS no Guia do usuário da interface de linha de comando.AWS

- Para obter detalhes da API, consulte CreateLaunchConfigurationem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo cria uma configuração de inicialização chamada “my-lc”. As instâncias do EC2 iniciadas por grupos do Auto Scaling que usam essa configuração de inicialização usam o tipo de instância, a AMI, o grupo de segurança e o perfil do IAM especificados.

```
New-ASLaunchConfiguration -LaunchConfigurationName my-lc -InstanceType
 "m3.medium" -ImageId "ami-12345678" -SecurityGroup "sg-12345678" -
IamInstanceProfile "myIamRole"
```

- Para obter detalhes da API, consulte CreateLaunchConfigurationem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo cria uma configuração de inicialização chamada “my-lc”. As instâncias do EC2 iniciadas por grupos do Auto Scaling que usam essa configuração de inicialização usam o tipo de instância, a AMI, o grupo de segurança e o perfil do IAM especificados.

```
New-ASLaunchConfiguration -LaunchConfigurationName my-lc -InstanceType
 "m3.medium" -ImageId "ami-12345678" -SecurityGroup "sg-12345678" -
IamInstanceProfile "myIamRole"
```

- Para obter detalhes da API, consulte CreateLaunchConfigurationem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`CreateOrUpdateTags`** com uma CLI

Os exemplos de código a seguir mostram como usar o `CreateOrUpdateTags`.

CLI

AWS CLI

Para criar ou modificar tags de um grupo do Auto Scaling existente

Este exemplo adiciona duas tags ao grupo do Auto Scaling especificado.

```
aws autoscaling create-or-update-tags \
    --tags ResourceId=my-asg,ResourceType=auto-scaling-
group,Key=Role,Value=WebServer,PropagateAtLaunch=true ResourceId=my-
asg,ResourceType=auto-scaling-
group,Key=Dept,Value=Research,PropagateAtLaunch=true
```

Este comando não produz saída.

Para obter mais informações, consulte Marcar instâncias e grupos de Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte CreateOrUpdateTagsem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo adiciona uma única tag ao grupo do Auto Scaling especificado. A chave da tag é 'myTag' e o valor da tag é 'myTagValue'. O Auto Scaling propaga essa tag para as instâncias do EC2 subsequentes lançadas pelo grupo do Auto Scaling. A sintaxe usada neste exemplo requer a PowerShell versão 3 ou posterior.

```
Set-ASTag -Tag @( @{ResourceType="auto-scaling-group"; ResourceId="my-asg";
 Key="myTag"; Value="myTagValue"; PropagateAtLaunch=$true} )
```

Exemplo 2: Com a PowerShell versão 2, você deve usar New-Object para criar a tag para o parâmetro Tag.

```
$tag = New-Object Amazon.AutoScaling.Model.Tag
$tag.ResourceType = "auto-scaling-group"
$tag.ResourceId = "my-asg"
$tag.Key = "myTag"
$tag.Value = "myTagValue"
$tag.PropagateAtLaunch = $true
Set-ASTag -Tag $tag
```

- Para obter detalhes da API, consulte CreateOrUpdateTagsem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo adiciona uma única tag ao grupo do Auto Scaling especificado. A chave da tag é 'myTag' e o valor da tag é 'myTagValue'. O Auto Scaling propaga essa tag para as instâncias do EC2 subsequentes lançadas pelo grupo do Auto Scaling. A sintaxe usada neste exemplo requer a PowerShell versão 3 ou posterior.

```
Set-ASTag -Tag @( @{ResourceType="auto-scaling-group"; ResourceId="my-asg";
 Key="myTag"; Value="myTagValue"; PropagateAtLaunch=$true} )
```

Exemplo 2: Com a PowerShell versão 2, você deve usar New-Object para criar a tag para o parâmetro Tag.

```
$tag = New-Object Amazon.AutoScaling.Model.Tag
$tag.ResourceType = "auto-scaling-group"
$tag.ResourceId = "my-asg"
$tag.Key = "myTag"
$tag.Value = "myTagValue"
$tag.PropagateAtLaunch = $true
Set-ASTag -Tag $tag
```

- Para obter detalhes da API, consulte CreateOrUpdateTagsem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Use **`DeleteAutoScalingGroup`** com um AWS SDK ou CLI

Os exemplos de código a seguir mostram como usar o `DeleteAutoScalingGroup`.

Exemplos de ações são trechos de código de programas maiores e devem ser executados em contexto. É possível ver essa ação em contexto nos seguintes exemplos de código:

- Conheça os conceitos básicos
- Criar e gerenciar um serviço resiliente

.NET

SDK para .NET

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

Atualize o tamanho mínimo de um grupo do Auto Scaling para zero, encerre todas as instâncias no grupo e exclua o grupo.

```
    /// <summary>
    /// Try to terminate an instance by its Id.
    /// </summary>
    /// <param name="instanceId">The Id of the instance to terminate.</param>
    /// <returns>Async task.</returns>
    public async Task TryTerminateInstanceById(string instanceId)
    {
        var stopping = false;
        Console.WriteLine($"Stopping {instanceId}...");
        while (!stopping)
        {
            try
            {
                await
_amazonAutoScaling.TerminateInstanceInAutoScalingGroupAsync(
                    new TerminateInstanceInAutoScalingGroupRequest()
                    {
                        InstanceId = instanceId,
                        ShouldDecrementDesiredCapacity = false
                    });
                stopping = true;
            }
            catch (ScalingActivityInProgressException)
            {
                Console.WriteLine($"Scaling activity in progress for
{instanceId}. Waiting...");
                Thread.Sleep(10000);
            }
        }
    }

    /// <summary>
    /// Tries to delete the EC2 Auto Scaling group. If the group is in use or in
progress,
    /// waits and retries until the group is successfully deleted.
    /// </summary>
    /// <param name="groupName">The name of the group to try to delete.</param>
    /// <returns>Async task.</returns>
    public async Task TryDeleteGroupByName(string groupName)
    {
```

```
        var stopped = false;
        while (!stopped)
        {
            try
            {
                await _amazonAutoScaling.DeleteAutoScalingGroupAsync(
                    new DeleteAutoScalingGroupRequest()
                    {
                        AutoScalingGroupName = groupName
                    });
                stopped = true;
            }
            catch (Exception e)
                when ((e is ScalingActivityInProgressException)
                      || (e is Amazon.AutoScaling.Model.ResourceInUseException))
            {
                Console.WriteLine($"Some instances are still running.
Waiting...");
                Thread.Sleep(10000);
            }
        }
    }

    /// <summary>
    /// Terminate instances and delete the Auto Scaling group by name.
    /// </summary>
    /// <param name="groupName">The name of the group to delete.</param>
    /// <returns>Async task.</returns>
    public async Task TerminateAndDeleteAutoScalingGroupWithName(string
groupName)
    {
        var describeGroupsResponse = await
_amazonAutoScaling.DescribeAutoScalingGroupsAsync(
            new DescribeAutoScalingGroupsRequest()
            {
                AutoScalingGroupNames = new List<string>() { groupName }
            });
        if (describeGroupsResponse.AutoScalingGroups.Any())
        {
            // Update the size to 0.
            await _amazonAutoScaling.UpdateAutoScalingGroupAsync(
                new UpdateAutoScalingGroupRequest()
                {
                    AutoScalingGroupName = groupName,
```

```
                    MinSize = 0
                });
            var group = describeGroupsResponse.AutoScalingGroups[0];
            foreach (var instance in group.Instances)
            {
                await TryTerminateInstanceById(instance.InstanceId);
            }

            await TryDeleteGroupByName(groupName);
        }
        else
        {
            Console.WriteLine($"No groups found with name {groupName}.");
        }
    }
```

```
    /// <summary>
    /// Delete an Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteAutoScalingGroupAsync(
        string groupName)
    {
        var deleteAutoScalingGroupRequest = new DeleteAutoScalingGroupRequest
        {
            AutoScalingGroupName = groupName,
            ForceDelete = true,
        };

        var response = await
 _amazonAutoScaling.DeleteAutoScalingGroupAsync(deleteAutoScalingGroupRequest);
        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
            Console.WriteLine($"You successfully deleted {groupName}");
            return true;
        }

        Console.WriteLine($"Couldn't delete {groupName}.");
        return false;
```

```
    }
```

- Para obter detalhes da API, consulte DeleteAutoScalingGroupa Referência AWS SDK para .NET da API.

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

            Aws::AutoScaling::Model::DeleteAutoScalingGroupRequest request;
            request.SetAutoScalingGroupName(groupName);

            Aws::AutoScaling::Model::DeleteAutoScalingGroupOutcome outcome =
                    autoScalingClient.DeleteAutoScalingGroup(request);

            if (outcome.IsSuccess()) {
                std::cout << "Auto Scaling group '" << groupName << "' was
deleted."
                          << std::endl;
            }
            else {
                std::cerr << "Error with AutoScaling::DeleteAutoScalingGroup. "
                          << outcome.GetError().GetMessage()
                          << std::endl;
                result = false;
            }
        }
```

- Para obter detalhes da API, consulte DeleteAutoScalingGroupa Referência AWS SDK para C++ da API.

CLI

AWS CLI

Exemplo 1: como excluir o grupo do Auto Scaling especificado

Este exemplo exclui o grupo do Auto Scaling especificado.

```
aws autoscaling delete-auto-scaling-group \
    --auto-scaling-group-name my-asg
```

Este comando não produz saída.

Para obter mais informações, consulte Excluir infraestrutura do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: como forçar a exclusão do grupo do Auto Scaling especificado

Use a opção `--force-delete` para excluir o grupo do Auto Scaling sem precisar esperar que as instâncias do grupo sejam encerradas.

```
aws autoscaling delete-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --force-delete
```

Este comando não produz saída.

Para obter mais informações, consulte Excluir infraestrutura do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DeleteAutoScalingGroupem Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.autoscaling.AutoScalingClient;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingException;
import
 software.amazon.awssdk.services.autoscaling.model.DeleteAutoScalingGroupRequest;

/**
 * Before running this SDK for Java (v2) code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class DeleteAutoScalingGroup {
    public static void main(String[] args) {
        final String usage = """

                Usage:
                    <groupName>

                Where:
                    groupName - The name of the Auto Scaling group.
                """;

        if (args.length != 1) {
            System.out.println(usage);
            System.exit(1);
        }

        String groupName = args[0];
```

```
        AutoScalingClient autoScalingClient = AutoScalingClient.builder()
                .region(Region.US_EAST_1)
                .build();

        deleteAutoScalingGroup(autoScalingClient, groupName);
        autoScalingClient.close();
    }

    public static void deleteAutoScalingGroup(AutoScalingClient
 autoScalingClient, String groupName) {
        try {
            DeleteAutoScalingGroupRequest deleteAutoScalingGroupRequest =
 DeleteAutoScalingGroupRequest.builder()
                    .autoScalingGroupName(groupName)
                    .forceDelete(true)
                    .build();


 autoScalingClient.deleteAutoScalingGroup(deleteAutoScalingGroupRequest);
            System.out.println("You successfully deleted " + groupName);

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- Para obter detalhes da API, consulte DeleteAutoScalingGroupa Referência AWS SDK for Java 2.x da API.

Kotlin

SDK para Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun deleteSpecificAutoScalingGroup(groupName: String) {
    val deleteAutoScalingGroupRequest =
        DeleteAutoScalingGroupRequest {
            autoScalingGroupName = groupName
            forceDelete = true
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.deleteAutoScalingGroup(deleteAutoScalingGroupRequest)
        println("You successfully deleted $groupName")
    }
}
```

- Para obter detalhes da API, consulte a DeleteAutoScalingGroupreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function deleteAutoScalingGroup($autoScalingGroupName)
    {
        return $this->autoScalingClient->deleteAutoScalingGroup([
            'AutoScalingGroupName' => $autoScalingGroupName,
            'ForceDelete' => true,
        ]);
    }
```

- Para obter detalhes da API, consulte DeleteAutoScalingGroupa Referência AWS SDK para PHP da API.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo exclui o grupo do Auto Scaling especificado se ele não tiver instâncias em execução. A confirmação será solicitada antes que a operação continue.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASAutoScalingGroup (DeleteAutoScalingGroup)" on
 Target "my-asg".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg -Force
```

Exemplo 3: Este exemplo exclui o grupo do Auto Scaling especificado e encerra todas as instâncias em execução que ele contém.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg -ForceDelete $true -Force
```

- Para obter detalhes da API, consulte DeleteAutoScalingGroupem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo exclui o grupo do Auto Scaling especificado se ele não tiver instâncias em execução. A confirmação será solicitada antes que a operação continue.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg
```

Saída:

```
Confirm
```

```
Are you sure you want to perform this action?
Performing operation "Remove-ASAutoScalingGroup (DeleteAutoScalingGroup)" on
 Target "my-asg".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg -Force
```

Exemplo 3: Este exemplo exclui o grupo do Auto Scaling especificado e encerra todas as instâncias em execução que ele contém.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg -ForceDelete $true -Force
```

- Para obter detalhes da API, consulte DeleteAutoScalingGroupem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

Atualize o tamanho mínimo de um grupo do Auto Scaling para zero, encerre todas as instâncias no grupo e exclua o grupo.

```
class AutoScalingWrapper:
    """
    Encapsulates Amazon EC2 Auto Scaling and EC2 management actions.
    """

    def __init__(
        self,
        resource_prefix: str,
```

```
        inst_type: str,
        ami_param: str,
        autoscaling_client: boto3.client,
        ec2_client: boto3.client,
        ssm_client: boto3.client,
        iam_client: boto3.client,
    ):
        """
        Initializes the AutoScaler class with the necessary parameters.

        :param resource_prefix: The prefix for naming AWS resources that are
created by this class.
        :param inst_type: The type of EC2 instance to create, such as t3.micro.
        :param ami_param: The Systems Manager parameter used to look up the AMI
that is created.
        :param autoscaling_client: A Boto3 EC2 Auto Scaling client.
        :param ec2_client: A Boto3 EC2 client.
        :param ssm_client: A Boto3 Systems Manager client.
        :param iam_client: A Boto3 IAM client.
        """
        self.inst_type = inst_type
        self.ami_param = ami_param
        self.autoscaling_client = autoscaling_client
        self.ec2_client = ec2_client
        self.ssm_client = ssm_client
        self.iam_client = iam_client
        sts_client = boto3.client("sts")
        self.account_id = sts_client.get_caller_identity()["Account"]

        self.key_pair_name = f"{resource_prefix}-key-pair"
        self.launch_template_name = f"{resource_prefix}-template-"
        self.group_name = f"{resource_prefix}-group"

        # Happy path
        self.instance_policy_name = f"{resource_prefix}-pol"
        self.instance_role_name = f"{resource_prefix}-role"
        self.instance_profile_name = f"{resource_prefix}-prof"

        # Failure mode
        self.bad_creds_policy_name = f"{resource_prefix}-bc-pol"
        self.bad_creds_role_name = f"{resource_prefix}-bc-role"
        self.bad_creds_profile_name = f"{resource_prefix}-bc-prof"
```

```
    def delete_autoscaling_group(self, group_name: str) -> None:
        """
        Terminates all instances in the group, then deletes the EC2 Auto Scaling
 group.

        :param group_name: The name of the group to delete.
        """
        try:
            response = self.autoscaling_client.describe_auto_scaling_groups(
                AutoScalingGroupNames=[group_name]
            )
            groups = response.get("AutoScalingGroups", [])
            if len(groups) > 0:
                self.autoscaling_client.update_auto_scaling_group(
                    AutoScalingGroupName=group_name, MinSize=0
                )
                instance_ids = [inst["InstanceId"] for inst in groups[0]
["Instances"]]
                for inst_id in instance_ids:
                    self.terminate_instance(inst_id)

                # Wait for all instances to be terminated
                if instance_ids:
                    waiter = self.ec2_client.get_waiter("instance_terminated")
                    log.info("Waiting for all instances to be terminated...")
                    waiter.wait(InstanceIds=instance_ids)
                    log.info("All instances have been terminated.")
            else:
                log.info(f"No groups found named '{group_name}'! Nothing to do.")
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(f"Failed to delete Auto Scaling group '{group_name}'.")
            if error_code == "ScalingActivityInProgressFault":
                log.error(
                    "Scaling activity is currently in progress. "
                    "Wait for the scaling activity to complete before attempting
 to delete the group again."
                )
            elif error_code == "ResourceContentionFault":
                log.error(
                    "The request failed due to a resource contention issue. "
                    "Ensure that no conflicting operations are being performed on
 the group."
                )
```

```
            log.error(f"Full error:\n\t{err}")
```

- Para obter detalhes da API, consulte a DeleteAutoScalingGroupReferência da API AWS SDK for Python (Boto3).

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
async fn delete_group(client: &Client, name: &str, force: bool) -> Result<(),
 Error> {
    client
        .delete_auto_scaling_group()
        .auto_scaling_group_name(name)
        .set_force_delete(if force { Some(true) } else { None })
        .send()
        .await?;

    println!("Deleted Auto Scaling group");

    Ok(())
}
```

- Para obter detalhes da API, consulte a DeleteAutoScalingGroupreferência da API AWS SDK for Rust.

SAP ABAP

SDK para SAP ABAP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    " Example: iv_group_name = 'my-auto-scaling-group'

    TRY.
        ao_asc->deleteautoscalinggroup(
          iv_autoscalinggroupname = iv_group_name ).

        " Wait for the group to be deleted (simplified - in production use proper
polling)
        WAIT UP TO 10 SECONDS.

        MESSAGE 'Auto Scaling group deleted successfully' TYPE 'I'.

      CATCH /aws1/cx_ascscaactivityinprg00 INTO DATA(lo_activity_in_progress).
        RAISE EXCEPTION lo_activity_in_progress.
      CATCH /aws1/cx_ascresourceinusefault INTO DATA(lo_resource_in_use).
        RAISE EXCEPTION lo_resource_in_use.
      CATCH /aws1/cx_rt_generic INTO DATA(lo_generic_exception).
        RAISE EXCEPTION lo_generic_exception.
    ENDTRY.
```

- Para obter detalhes da API, consulte a DeleteAutoScalingGroupreferência da API AWS SDK for SAP ABAP.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DeleteLaunchConfiguration`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DeleteLaunchConfiguration`.

CLI

AWS CLI

Para excluir sua configuração de ativação

Este exemplo exclui a configuração de execução especificada.

```
aws autoscaling delete-launch-configuration \
    --launch-configuration-name my-launch-config
```

Este comando não produz saída.

Para obter mais informações, consulte Excluir infraestrutura do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DeleteLaunchConfigurationem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo exclui a configuração de inicialização especificada se ela não estiver anexada a um grupo do Auto Scaling. A confirmação será solicitada antes que a operação continue.

```
Remove-ASLaunchConfiguration -LaunchConfigurationName my-lc
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASLaunchConfiguration (DeleteLaunchConfiguration)"
 on Target "my-lc".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASLaunchConfiguration -LaunchConfigurationName my-lc -Force
```

- Para obter detalhes da API, consulte DeleteLaunchConfigurationem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo exclui a configuração de inicialização especificada se ela não estiver anexada a um grupo do Auto Scaling. A confirmação será solicitada antes que a operação continue.

```
Remove-ASLaunchConfiguration -LaunchConfigurationName my-lc
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASLaunchConfiguration (DeleteLaunchConfiguration)"
 on Target "my-lc".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASLaunchConfiguration -LaunchConfigurationName my-lc -Force
```

- Para obter detalhes da API, consulte DeleteLaunchConfigurationem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DeleteLifecycleHook`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DeleteLifecycleHook`.

CLI

AWS CLI

Para adicionar um gancho do ciclo de vida.

Este exemplo exclui o gancho do ciclo de vida especificado.

```
aws autoscaling delete-lifecycle-hook \
    --lifecycle-hook-name my-lifecycle-hook \
    --auto-scaling-group-name my-asg
```

Este comando não produz saída.

- Para obter detalhes da API, consulte DeleteLifecycleHookem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: esse exemplo exclui o gancho do ciclo de vida especificado para o grupo do Auto Scaling especificado. A confirmação será solicitada antes que a operação continue.

```
Remove-ASLifecycleHook -AutoScalingGroupName my-asg -LifecycleHookName
 myLifecycleHook
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASLifecycleHook (DeleteLifecycleHook)" on Target
 "myLifecycleHook".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASLifecycleHook -AutoScalingGroupName my-asg -LifecycleHookName
 myLifecycleHook -Force
```

- Para obter detalhes da API, consulte DeleteLifecycleHookem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: esse exemplo exclui o gancho do ciclo de vida especificado para o grupo do Auto Scaling especificado. A confirmação será solicitada antes que a operação continue.

```
Remove-ASLifecycleHook -AutoScalingGroupName my-asg -LifecycleHookName
 myLifecycleHook
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASLifecycleHook (DeleteLifecycleHook)" on Target
 "myLifecycleHook".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASLifecycleHook -AutoScalingGroupName my-asg -LifecycleHookName
 myLifecycleHook -Force
```

- Para obter detalhes da API, consulte DeleteLifecycleHookem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DeleteNotificationConfiguration`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DeleteNotificationConfiguration`.

CLI

AWS CLI

Como excluir uma notificação do Auto Scaling

Este exemplo exclui a notificação especificada do grupo do Auto Scaling especificado.

```
aws autoscaling delete-notification-configuration \
    --auto-scaling-group-name my-asg \
    --topic-arn arn:aws:sns:us-west-2:123456789012:my-sns-topic
```

Este comando não produz saída.

Para obter mais informações, consulte Excluir a configuração de notificação no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DeleteNotificationConfigurationem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo exclui a ação de notificação especificada. A confirmação será solicitada antes que a operação continue.

```
Remove-ASNotificationConfiguration -AutoScalingGroupName my-asg -TopicARN
 "arn:aws:sns:us-west-2:123456789012:my-topic"
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASNotificationConfiguration
 (DeleteNotificationConfiguration)" on Target
"arn:aws:sns:us-west-2:123456789012:my-topic".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASNotificationConfiguration -AutoScalingGroupName my-asg -TopicARN
 "arn:aws:sns:us-west-2:123456789012:my-topic" -Force
```

- Para obter detalhes da API, consulte DeleteNotificationConfigurationem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo exclui a ação de notificação especificada. A confirmação será solicitada antes que a operação continue.

```
Remove-ASNotificationConfiguration -AutoScalingGroupName my-asg -TopicARN
 "arn:aws:sns:us-west-2:123456789012:my-topic"
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASNotificationConfiguration
 (DeleteNotificationConfiguration)" on Target
"arn:aws:sns:us-west-2:123456789012:my-topic".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASNotificationConfiguration -AutoScalingGroupName my-asg -TopicARN
 "arn:aws:sns:us-west-2:123456789012:my-topic" -Force
```

- Para obter detalhes da API, consulte DeleteNotificationConfigurationem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DeletePolicy`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DeletePolicy`.

CLI

AWS CLI

Excluir uma política de escalabilidade

Este exemplo exclui a política de escalabilidade especificada.

```
aws autoscaling delete-policy \
    --auto-scaling-group-name my-asg \
    --policy-name alb1000-target-tracking-scaling-policy
```

Este comando não produz saída.

- Para obter detalhes da API, consulte DeletePolicyem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo exclui a política especificada para o grupo do Auto Scaling especificado. A confirmação será solicitada antes que a operação continue.

```
Remove-ASPolicy -AutoScalingGroupName my-asg -PolicyName myScaleInPolicy
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASPolicy (DeletePolicy)" on Target
 "myScaleInPolicy".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASPolicy -AutoScalingGroupName my-asg -PolicyName myScaleInPolicy -Force
```

- Para obter detalhes da API, consulte DeletePolicyem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo exclui a política especificada para o grupo do Auto Scaling especificado. A confirmação será solicitada antes que a operação continue.

```
Remove-ASPolicy -AutoScalingGroupName my-asg -PolicyName myScaleInPolicy
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASPolicy (DeletePolicy)" on Target
 "myScaleInPolicy".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASPolicy -AutoScalingGroupName my-asg -PolicyName myScaleInPolicy -Force
```

- Para obter detalhes da API, consulte DeletePolicyem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DeleteScheduledAction`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DeleteScheduledAction`.

CLI

AWS CLI

Para criar uma ação programada para um grupo do Auto Scaling

Este exemplo exclui a ação programada especificada do grupo do Auto Scaling especificado.

```
aws autoscaling delete-scheduled-action \
    --auto-scaling-group-name my-asg \
    --scheduled-action-name my-scheduled-action
```

Este comando não produz saída.

- Para obter detalhes da API, consulte DeleteScheduledActionem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo exclui a política especificada para o grupo do Auto Scaling especificado. A confirmação será solicitada antes que a operação continue.

```
Remove-ASScheduledAction -AutoScalingGroupName my-asg -ScheduledAction
 "myScheduledAction"
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASScheduledAction (DeleteScheduledAction)" on Target
 "myScheduledAction".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASScheduledAction -AutoScalingGroupName my-asg -ScheduledAction
 "myScheduledAction" -Force
```

- Para obter detalhes da API, consulte DeleteScheduledActionem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo exclui a política especificada para o grupo do Auto Scaling especificado. A confirmação será solicitada antes que a operação continue.

```
Remove-ASScheduledAction -AutoScalingGroupName my-asg -ScheduledAction
 "myScheduledAction"
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASScheduledAction (DeleteScheduledAction)" on Target
 "myScheduledAction".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASScheduledAction -AutoScalingGroupName my-asg -ScheduledAction
 "myScheduledAction" -Force
```

- Para obter detalhes da API, consulte DeleteScheduledActionem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DeleteTags`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DeleteTags`.

CLI

AWS CLI

Como excluir uma tag de um grupo do Auto Scaling

Este exemplo exclui a tag especificada do grupo do Auto Scaling especificado.

```
aws autoscaling delete-tags \
    --tags ResourceId=my-asg,ResourceType=auto-scaling-
group,Key=Dept,Value=Research
```

Este comando não produz saída.

Para obter mais informações, consulte Marcar instâncias e grupos de Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DeleteTagsem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: esse exemplo remove a tag especificada do grupo do Auto Scaling especificado. A confirmação será solicitada antes que a operação continue. A sintaxe usada neste exemplo requer a PowerShell versão 3 ou posterior.

```
Remove-ASTag -Tag @( @{ResourceType="auto-scaling-group"; ResourceId="my-asg";
 Key="myTag" } )
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing the operation "Remove-ASTag (DeleteTags)" on target
 "Amazon.AutoScaling.Model.Tag".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASTag -Tag @( @{ResourceType="auto-scaling-group"; ResourceId="my-asg";
 Key="myTag" } ) -Force
```

Exemplo 3: com o PowerShell versão, é necessário usar New-Object para criar a tag para o parâmetro de Tag.

```
$tag = New-Object Amazon.AutoScaling.Model.Tag
$tag.ResourceType = "auto-scaling-group"
$tag.ResourceId = "my-asg"
$tag.Key = "myTag"
```

```
Remove-ASTag -Tag $tag -Force
```

- Para obter detalhes da API, consulte DeleteTagsem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: esse exemplo remove a tag especificada do grupo do Auto Scaling especificado. A confirmação será solicitada antes que a operação continue. A sintaxe usada neste exemplo requer a PowerShell versão 3 ou posterior.

```
Remove-ASTag -Tag @( @{ResourceType="auto-scaling-group"; ResourceId="my-asg";
 Key="myTag" } )
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing the operation "Remove-ASTag (DeleteTags)" on target
 "Amazon.AutoScaling.Model.Tag".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes de prosseguir com a operação.

```
Remove-ASTag -Tag @( @{ResourceType="auto-scaling-group"; ResourceId="my-asg";
 Key="myTag" } ) -Force
```

Exemplo 3: com o PowerShell versão, é necessário usar New-Object para criar a tag para o parâmetro de Tag.

```
$tag = New-Object Amazon.AutoScaling.Model.Tag
$tag.ResourceType = "auto-scaling-group"
$tag.ResourceId = "my-asg"
$tag.Key = "myTag"
Remove-ASTag -Tag $tag -Force
```

- Para obter detalhes da API, consulte DeleteTagsem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar `DescribeAccountLimits` com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeAccountLimits`.

CLI

AWS CLI

Como descrever os limites de sua conta do Amazon EC2 Auto Scaling

Este exemplo descreve os limites AWS do Amazon EC2 Auto Scaling para sua conta.

```
aws autoscaling describe-account-limits
```

Saída:

```
{
    "NumberOfLaunchConfigurations": 5,
    "MaxNumberOfLaunchConfigurations": 100,
    "NumberOfAutoScalingGroups": 3,
    "MaxNumberOfAutoScalingGroups": 20
}
```

Para obter mais informações, consulte Service quotas do Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DescribeAccountLimitsem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo descreve os limites de recursos do Auto Scaling para sua AWS conta.

```
Get-ASAccountLimit
```

Saída:

```
MaxNumberOfAutoScalingGroups    : 20
MaxNumberOfLaunchConfigurations : 100
```

- Para obter detalhes da API, consulte DescribeAccountLimitsem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo descreve os limites de recursos do Auto Scaling para sua AWS conta.

```
Get-ASAccountLimit
```

Saída:

```
MaxNumberOfAutoScalingGroups    : 20
MaxNumberOfLaunchConfigurations : 100
```

- Para obter detalhes da API, consulte DescribeAccountLimitsem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribeAdjustmentTypes`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeAdjustmentTypes`.

CLI

AWS CLI

Como descrever os tipos de ajuste de escala disponíveis

Este exemplo descreve os tipos de ajuste disponíveis.

```
aws autoscaling describe-adjustment-types
```

Saída:

```
{
    "AdjustmentTypes": [
        {
            "AdjustmentType": "ChangeInCapacity"
        },
        {
            "AdjustmentType": "ExactCapacity"
        },
        {
            "AdjustmentType": "PercentChangeInCapacity"
        }
    ]
}
```

Para obter mais informações, consulte Tipos de ajuste da escalabilidade no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DescribeAdjustmentTypesem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo descreve os tipos de ajuste que são compatíveis com o Auto Scaling.

```
Get-ASAdjustmentType
```

Saída:

```
Type
----
ChangeInCapacity
ExactCapacity
PercentChangeInCapacity
```

- Para obter detalhes da API, consulte DescribeAdjustmentTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo descreve os tipos de ajuste que são compatíveis com o Auto Scaling.

```
Get-ASAdjustmentType
```

Saída:

```
Type
----
ChangeInCapacity
ExactCapacity
PercentChangeInCapacity
```

- Para obter detalhes da API, consulte DescribeAdjustmentTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Use **`DescribeAutoScalingGroups`** com um AWS SDK ou CLI

Os exemplos de código a seguir mostram como usar o `DescribeAutoScalingGroups`.

Exemplos de ações são trechos de código de programas maiores e devem ser executados em contexto. É possível ver essa ação em contexto nos seguintes exemplos de código:

- Conheça os conceitos básicos
- Criar e gerenciar um serviço resiliente

.NET

SDK para .NET (v4)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    /// <summary>
    /// Get data about the instances in an Amazon EC2 Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A list of Amazon EC2 Auto Scaling details.</returns>
    public async Task<List<AutoScalingInstanceDetails>>
 DescribeAutoScalingInstancesAsync(
        string groupName)
    {
        var groups = await DescribeAutoScalingGroupsAsync(groupName);
        var instanceIds = new List<string>();
        var instanceDetails = new List<AutoScalingInstanceDetails>();
        if (groups != null)
        {
            groups.ForEach(group =>
            {
                if (group.AutoScalingGroupName == groupName && group.Instances !=
 null)
                {
                    group.Instances.ForEach(instance =>
                    {
                        instanceIds.Add(instance.InstanceId);
                    });
                }
            });

            var scalingGroupsRequest = new DescribeAutoScalingInstancesRequest
            {
                MaxRecords = 10,
                InstanceIds = instanceIds,
```

```
            };

            var response =
                await _amazonAutoScaling.DescribeAutoScalingInstancesAsync(
                    scalingGroupsRequest);
            if (response.AutoScalingInstances != null)
            {
                instanceDetails = response.AutoScalingInstances;
            }
        }

        return instanceDetails;
    }
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK para .NET da API.

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    Aws::Client::ClientConfiguration clientConfig;
    // Optional: Set to the AWS Region (overrides config file).
    // clientConfig.region = "us-east-1";

Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

Aws::AutoScaling::Model::DescribeAutoScalingGroupsRequest request;
Aws::Vector<Aws::String> groupNames;
groupNames.push_back(groupName);
request.SetAutoScalingGroupNames(groupNames);

Aws::AutoScaling::Model::DescribeAutoScalingGroupsOutcome outcome =
```

```
            client.DescribeAutoScalingGroups(request);

    if (outcome.IsSuccess()) {
        autoScalingGroup = outcome.GetResult().GetAutoScalingGroups();
    }
    else {
        std::cerr << "Error with AutoScaling::DescribeAutoScalingGroups. "
                  << outcome.GetError().GetMessage()
                  << std::endl;
    }
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK para C++ da API.

CLI

AWS CLI

Exemplo 1: como descrever o grupo do Auto Scaling especificado

Este exemplo descreve o grupo do Auto Scaling especificado.

```
aws autoscaling describe-auto-scaling-groups \
    --auto-scaling-group-names my-asg
```

Saída:

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:930d940e-891e-4781-
a11a-7b0acd480f03:autoScalingGroupName/my-asg",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-1234567890abcde12"
            },
            "MinSize": 0,
            "MaxSize": 1,
```

```
            "DesiredCapacity": 1,
            "DefaultCooldown": 300,
            "AvailabilityZones": [
                "us-west-2a",
                "us-west-2b",
                "us-west-2c"
            ],
            "LoadBalancerNames": [],
            "TargetGroupARNs": [],
            "HealthCheckType": "EC2",
            "HealthCheckGracePeriod": 0,
            "Instances": [
                {
                    "InstanceId": "i-06905f55584de02da",
                    "InstanceType": "t2.micro",
                    "AvailabilityZone": "us-west-2a",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService",
                    "ProtectedFromScaleIn": false,
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-1234567890abcde12"
                    }
                }
            ],
            "CreatedTime": "2023-10-28T02:39:22.152Z",
            "SuspendedProcesses": [],
            "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-
c934b782",
            "EnabledMetrics": [],
            "Tags": [],
            "TerminationPolicies": [
                "Default"
            ],
            "NewInstancesProtectedFromScaleIn": false,
            "ServiceLinkedRoleARN":"arn",
            "TrafficSources": []
        }
    ]
}
```

Exemplo 2: como descrever os primeiros 100 grupos do Auto Scaling especificados

Este exemplo descreve os grupos do Auto Scaling especificados. Ele permite especificar até cem nomes de grupos.

```
aws autoscaling describe-auto-scaling-groups \
    --max-items 100 \
    --auto-scaling-group-names "group1" "group2" "group3" "group4"
```

Consulte uma amostra de saída no exemplo 1.

Exemplo 3: como descrever um grupo do Auto Scaling na região especificada

Este exemplo descreve até 75 grupos do Auto Scaling na região especificada.

```
aws autoscaling describe-auto-scaling-groups \
    --max-items 75 \
    --region us-east-1
```

Consulte um exemplo de saída no exemplo 1.

Exemplo 4: como descrever o número especificado do grupo do Auto Scaling

Use a opção `--max-items` para retornar um número específico de grupos do Auto Scaling.

```
aws autoscaling describe-auto-scaling-groups \
    --max-items 1
```

Consulte um exemplo de saída no exemplo 1.

Se a saída incluir um campo `NextToken`, há mais grupos. Para obter os grupos adicionais, use o valor desse campo com a opção `--starting-token` em uma chamada subsequente da seguinte maneira.

```
aws autoscaling describe-auto-scaling-groups \
    --starting-token Z3M3LMPEXAMPLE
```

Consulte um exemplo de saída no exemplo 1.

Exemplo 5: descrever grupos do Auto Scaling que usam configurações de execução

Este exemplo usa a opção `--query` para descrever os grupos do Auto Scaling que usam configurações de execução.

```
aws autoscaling describe-auto-scaling-groups \
    --query 'AutoScalingGroups[?LaunchConfigurationName!=`null`]'
```

Saída:

```
[
    {
        "AutoScalingGroupName": "my-asg",
        "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:930d940e-891e-4781-
a11a-7b0acd480f03:autoScalingGroupName/my-asg",
        "LaunchConfigurationName": "my-lc",
        "MinSize": 0,
        "MaxSize": 1,
        "DesiredCapacity": 1,
        "DefaultCooldown": 300,
        "AvailabilityZones": [
            "us-west-2a",
            "us-west-2b",
            "us-west-2c"
        ],
        "LoadBalancerNames": [],
        "TargetGroupARNs": [],
        "HealthCheckType": "EC2",
        "HealthCheckGracePeriod": 0,
        "Instances": [
            {
                "InstanceId": "i-088c57934a6449037",
                "InstanceType": "t2.micro",
                "AvailabilityZone": "us-west-2c",
                "HealthStatus": "Healthy",
                "LifecycleState": "InService",
                "LaunchConfigurationName": "my-lc",
                "ProtectedFromScaleIn": false
            }
        ],
        "CreatedTime": "2023-10-28T02:39:22.152Z",
        "SuspendedProcesses": [],
        "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
        "EnabledMetrics": [],
        "Tags": [],
        "TerminationPolicies": [
            "Default"
```

```
        ],
        "NewInstancesProtectedFromScaleIn": false,
        "ServiceLinkedRoleARN":"arn",
        "TrafficSources": []
    }
]
```

Para obter mais informações, consulte a saída da AWS CLI do filtro no Guia do usuário da interface de linha de AWS comando.

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsem Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.autoscaling.AutoScalingClient;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingException;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingGroup;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsResponse;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsRequest;
import software.amazon.awssdk.services.autoscaling.model.Instance;
import java.util.List;

/**
 * Before running this SDK for Java (v2) code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
```

```
 */
public class DescribeAutoScalingInstances {
    public static void main(String[] args) {
        final String usage = """

                Usage:
                    <groupName>

                Where:
                    groupName - The name of the Auto Scaling group.
                """;

        if (args.length != 1) {
            System.out.println(usage);
            System.exit(1);
        }

        String groupName = args[0];
        AutoScalingClient autoScalingClient = AutoScalingClient.builder()
                .region(Region.US_EAST_1)
                .build();

        String instanceId = getAutoScaling(autoScalingClient, groupName);
        System.out.println(instanceId);
        autoScalingClient.close();
    }

    public static String getAutoScaling(AutoScalingClient autoScalingClient,
 String groupName) {
        try {
            String instanceId = "";
            DescribeAutoScalingGroupsRequest scalingGroupsRequest =
 DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();

            DescribeAutoScalingGroupsResponse response = autoScalingClient
                    .describeAutoScalingGroups(scalingGroupsRequest);
            List<AutoScalingGroup> groups = response.autoScalingGroups();
            for (AutoScalingGroup group : groups) {
                System.out.println("The group name is " +
 group.autoScalingGroupName());
                System.out.println("The group ARN is " +
 group.autoScalingGroupARN());
```

```
                List<Instance> instances = group.instances();
                for (Instance instance : instances) {
                    instanceId = instance.instanceId();
                }
            }
            return instanceId;
        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
        return "";
    }
}
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK for Java 2.x da API.

Kotlin

SDK para Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun getAutoScalingGroups(groupName: String) {
    val scalingGroupsRequest =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        val response =
 autoScalingClient.describeAutoScalingGroups(scalingGroupsRequest)
        response.autoScalingGroups?.forEach { group ->
            println("The group name is ${group.autoScalingGroupName}")
            println("The group ARN is ${group.autoScalingGroupArn}")
```

```
            group.instances?.forEach { instance ->
                println("The instance id is ${instance.instanceId}")
                println("The lifecycle state is " + instance.lifecycleState)
            }
        }
    }
}
```

- Para obter detalhes da API, consulte a DescribeAutoScalingGroupsreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function describeAutoScalingGroups($autoScalingGroupNames)
    {
        return $this->autoScalingClient->describeAutoScalingGroups([
            'AutoScalingGroupNames' => $autoScalingGroupNames
        ]);
    }
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK para PHP da API.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo lista os nomes de seus grupos do Auto Scaling.

```
Get-ASAutoScalingGroup | format-table -property AutoScalingGroupName
```

Saída:

```
AutoScalingGroupName
--------------------
my-asg-1
my-asg-2
my-asg-3
my-asg-4
my-asg-5
my-asg-6
```

Exemplo 2: Este exemplo descreve o grupo do Auto Scaling especificado.

```
Get-ASAutoScalingGroup -AutoScalingGroupName my-asg-1
```

Saída:

```
AutoScalingGroupARN     : arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:930d940e-891e-4781-a11a-7b0acd480
                          f03:autoScalingGroupName/my-asg-1
AutoScalingGroupName    : my-asg-1
AvailabilityZones       : {us-west-2b, us-west-2a}
CreatedTime             : 3/1/2015 9:05:31 AM
DefaultCooldown         : 300
DesiredCapacity         : 2
EnabledMetrics          : {}
HealthCheckGracePeriod  : 300
HealthCheckType         : EC2
Instances               : {my-lc}
LaunchConfigurationName : my-lc
LoadBalancerNames       : {}
MaxSize                 : 0
MinSize                 : 0
PlacementGroup          :
Status                  :
SuspendedProcesses      : {}
Tags                    : {}
TerminationPolicies     : {Default}
VPCZoneIdentifier       : subnet-e4f33493,subnet-5264e837
```

Exemplo 3: Este exemplo descreve os grupos do Auto Scaling especificados.

```
Get-ASAutoScalingGroup -AutoScalingGroupName @("my-asg-1", "my-asg-2")
```

Exemplo 4: Este exemplo descreve as instâncias do Auto Scaling do grupo do Auto Scaling especificado.

```
(Get-ASAutoScalingGroup -AutoScalingGroupName my-asg-1).Instances
```

Exemplo 5: Este exemplo descreve todos os seus grupos do Auto Scaling.

```
Get-ASAutoScalingGroup
```

Exemplo 6: Este LaunchTemplate exemplo descreve o grupo de Auto Scaling especificado. Este exemplo pressupõe que as “Opções de compra de instância” estejam definidas como “Aderir ao modelo de lançamento”. Caso essa opção esteja definida como “Combinar opções de compra e tipos de instância”, LaunchTemplate pode ser acessada usando "MixedInstancesPolicy. LaunchTemplate” propriedade.

```
(Get-ASAutoScalingGroup -AutoScalingGroupName my-ag-1).LaunchTemplate
```

Saída:

```
LaunchTemplateId     LaunchTemplateName   Version
----------------     ------------------   -------
lt-06095fd619cb40371 test-launch-template $Default
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo lista os nomes de seus grupos do Auto Scaling.

```
Get-ASAutoScalingGroup | format-table -property AutoScalingGroupName
```

Saída:

```
AutoScalingGroupName
--------------------
my-asg-1
```

```
my-asg-2
my-asg-3
my-asg-4
my-asg-5
my-asg-6
```

Exemplo 2: Este exemplo descreve o grupo do Auto Scaling especificado.

```
Get-ASAutoScalingGroup -AutoScalingGroupName my-asg-1
```

Saída:

```
AutoScalingGroupARN     : arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:930d940e-891e-4781-a11a-7b0acd480
                          f03:autoScalingGroupName/my-asg-1
AutoScalingGroupName    : my-asg-1
AvailabilityZones       : {us-west-2b, us-west-2a}
CreatedTime             : 3/1/2015 9:05:31 AM
DefaultCooldown         : 300
DesiredCapacity         : 2
EnabledMetrics          : {}
HealthCheckGracePeriod  : 300
HealthCheckType         : EC2
Instances               : {my-lc}
LaunchConfigurationName : my-lc
LoadBalancerNames       : {}
MaxSize                 : 0
MinSize                 : 0
PlacementGroup          :
Status                  :
SuspendedProcesses      : {}
Tags                    : {}
TerminationPolicies     : {Default}
VPCZoneIdentifier       : subnet-e4f33493,subnet-5264e837
```

Exemplo 3: Este exemplo descreve os grupos do Auto Scaling especificados.

```
Get-ASAutoScalingGroup -AutoScalingGroupName @("my-asg-1", "my-asg-2")
```

Exemplo 4: Este exemplo descreve as instâncias do Auto Scaling do grupo do Auto Scaling especificado.

```
(Get-ASAutoScalingGroup -AutoScalingGroupName my-asg-1).Instances
```

Exemplo 5: Este exemplo descreve todos os seus grupos do Auto Scaling.

```
Get-ASAutoScalingGroup
```

Exemplo 6: Este LaunchTemplate exemplo descreve o grupo de Auto Scaling especificado. Este exemplo pressupõe que as “Opções de compra de instância” estejam definidas como “Aderir ao modelo de lançamento”. Caso essa opção esteja definida como “Combinar opções de compra e tipos de instância”, LaunchTemplate pode ser acessada usando "MixedInstancesPolicy. LaunchTemplate” propriedade.

```
(Get-ASAutoScalingGroup -AutoScalingGroupName my-ag-1).LaunchTemplate
```

Saída:

```
LaunchTemplateId     LaunchTemplateName   Version
----------------     ------------------   -------
lt-06095fd619cb40371 test-launch-template $Default
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
```

```
        """
        self.autoscaling_client = autoscaling_client


    def describe_group(self, group_name: str) -> Optional[Dict[str, Any]]:
        """
        Gets information about an Auto Scaling group.

        :param group_name: The name of the group to look up.
        :return: A dictionary with information about the group if found,
otherwise None.
        :raises ClientError: If there is an error describing the Auto Scaling
group.
        """
        try:
            paginator = self.autoscaling_client.get_paginator(
                "describe_auto_scaling_groups"
            )
            response_iterator =
paginator.paginate(AutoScalingGroupNames=[group_name])
            groups = []
            for response in response_iterator:
                groups.extend(response.get("AutoScalingGroups", []))

            logger.info(
                f"Successfully retrieved information for Auto Scaling group
{group_name}."
            )

        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            logger.error(f"Failed to describe Auto Scaling group {group_name}.")
            if error_code == "ResourceContentionFault":
                logger.error(
                    "There is a conflict with another operation that is modifying
the "
                    f"Auto Scaling group '{group_name}' Please try again later."
                )
            logger.error(f"Full error:\n\t{err}")
            raise
        else:
            return groups[0] if len(groups) > 0 else None
```

- Para obter detalhes da API, consulte a DescribeAutoScalingGroupsReferência da API AWS SDK for Python (Boto3).

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
async fn list_groups(client: &Client) -> Result<(), Error> {
    let resp = client.describe_auto_scaling_groups().send().await?;

    println!("Groups:");

    let groups = resp.auto_scaling_groups();

    for group in groups {
        println!(
            "Name:  {}",
            group.auto_scaling_group_name().unwrap_or("Unknown")
        );
        println!(
            "Arn:   {}",
            group.auto_scaling_group_arn().unwrap_or("unknown"),
        );
        println!("Zones: {:?}", group.availability_zones(),);
        println!();
    }

    println!("Found {} group(s)", groups.len());

    Ok(())
}
```

- Para obter detalhes da API, consulte a DescribeAutoScalingGroupsreferência da API AWS SDK for Rust.

SAP ABAP

SDK para SAP ABAP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    DATA lt_group_names TYPE /aws1/
cl_ascautoscgroupnames_w=>tt_autoscalinggroupnames.
    DATA lo_group_name TYPE REF TO /aws1/cl_ascautoscgroupnames_w.

    " Example: iv_group_name = 'my-auto-scaling-group'

    TRY.
        " Build group names parameter
        CREATE OBJECT lo_group_name
          EXPORTING
            iv_value = iv_group_name.
        APPEND lo_group_name TO lt_group_names.

        " Describe the Auto Scaling group
        DATA(lo_output) = ao_asc->describeautoscalinggroups(
          it_autoscalinggroupnames = lt_group_names ).

        " Return the first (and only) group in the result
        DATA(lt_groups) = lo_output->get_autoscalinggroups( ).
        IF lines( lt_groups ) > 0.
          READ TABLE lt_groups INDEX 1 INTO DATA(lo_group).
          oo_output = lo_group.
        ENDIF.

        MESSAGE 'Auto Scaling group information retrieved successfully' TYPE 'I'.

      CATCH /aws1/cx_ascresrccontionfault INTO DATA(lo_contention).
        RAISE EXCEPTION lo_contention.
```

```
      CATCH /aws1/cx_rt_generic INTO DATA(lo_generic_exception).
        RAISE EXCEPTION lo_generic_exception.
    ENDTRY.
```

- Para obter detalhes da API, consulte a DescribeAutoScalingGroupsreferência da API AWS SDK for SAP ABAP.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Use **`DescribeAutoScalingInstances`** com um AWS SDK ou CLI

Os exemplos de código a seguir mostram como usar o `DescribeAutoScalingInstances`.

Exemplos de ações são trechos de código de programas maiores e devem ser executados em contexto. É possível ver essa ação em contexto no seguinte exemplo de código:

- Conheça os conceitos básicos

.NET

SDK para .NET (v4)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    /// <summary>
    /// Get data about the instances in an Amazon EC2 Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A list of Amazon EC2 Auto Scaling details.</returns>
    public async Task<List<AutoScalingInstanceDetails>>
 DescribeAutoScalingInstancesAsync(
```

```
        string groupName)
    {
        var groups = await DescribeAutoScalingGroupsAsync(groupName);
        var instanceIds = new List<string>();
        var instanceDetails = new List<AutoScalingInstanceDetails>();
        if (groups != null)
        {
            groups.ForEach(group =>
            {
                if (group.AutoScalingGroupName == groupName && group.Instances !=
null)
                {
                    group.Instances.ForEach(instance =>
                    {
                        instanceIds.Add(instance.InstanceId);
                    });
                }
            });

            var scalingGroupsRequest = new DescribeAutoScalingInstancesRequest
            {
                MaxRecords = 10,
                InstanceIds = instanceIds,
            };

            var response =
                await _amazonAutoScaling.DescribeAutoScalingInstancesAsync(
                    scalingGroupsRequest);
            if (response.AutoScalingInstances != null)
            {
                instanceDetails = response.AutoScalingInstances;
            }
        }

        return instanceDetails;
    }
```

- Para obter detalhes da API, consulte DescribeAutoScalingInstancesa Referência AWS SDK para .NET da API.

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

        Aws::AutoScaling::Model::DescribeAutoScalingInstancesRequest request;
        request.SetInstanceIds(instanceIDs);

        Aws::AutoScaling::Model::DescribeAutoScalingInstancesOutcome outcome =
                client.DescribeAutoScalingInstances(request);

        if (outcome.IsSuccess()) {
            const
Aws::Vector<Aws::AutoScaling::Model::AutoScalingInstanceDetails>
&instancesDetails =
                    outcome.GetResult().GetAutoScalingInstances();

        }
        else {
            std::cerr << "Error with AutoScaling::DescribeAutoScalingInstances. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
            return false;
        }
```

- Para obter detalhes da API, consulte DescribeAutoScalingInstancesa Referência AWS SDK para C++ da API.

CLI

AWS CLI

Exemplo 1: como descrever uma ou mais instâncias

Este exemplo descreve a instância especificada.

```
aws autoscaling describe-auto-scaling-instances \
    --instance-ids i-06905f55584de02da
```

Saída:

```
{
    "AutoScalingInstances": [
        {
            "InstanceId": "i-06905f55584de02da",
            "InstanceType": "t2.micro",
            "AutoScalingGroupName": "my-asg",
            "AvailabilityZone": "us-west-2b",
            "LifecycleState": "InService",
            "HealthStatus": "HEALTHY",
            "ProtectedFromScaleIn": false,
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-1234567890abcde12",
                "LaunchTemplateName": "my-launch-template",
                "Version": "1"
            }
        }
    ]
}
```

Exemplo 2: como descrever uma ou mais instâncias

Este exemplo usa a opção `--max-items` para especificar quantas instâncias devem ser retornadas pela chamada.

```
aws autoscaling describe-auto-scaling-instances \
    --max-items 1
```

Se a saída incluir um campo `NextToken`, há mais instâncias. Para obter as instâncias adicionais, use o valor desse campo com a opção `--starting-token` em uma chamada subsequente da seguinte maneira.

```
aws autoscaling describe-auto-scaling-instances \
    --starting-token Z3M3LMPEXAMPLE
```

Consulte uma amostra de saída no exemplo 1.

Exemplo 3: descrever instâncias que usam configurações de execução

Este exemplo usa a opção `--query` para descrever instâncias que usam configurações de execução.

```
aws autoscaling describe-auto-scaling-instances \
    --query 'AutoScalingInstances[?LaunchConfigurationName!=`null`]'
```

Saída:

```
[
    {
        "InstanceId": "i-088c57934a6449037",
        "InstanceType": "t2.micro",
        "AutoScalingGroupName": "my-asg",
        "AvailabilityZone": "us-west-2c",
        "LifecycleState": "InService",
        "HealthStatus": "HEALTHY",
        "LaunchConfigurationName": "my-lc",
        "ProtectedFromScaleIn": false
    }
]
```

Para obter mais informações, consulte a saída da AWS CLI do filtro no Guia do usuário da interface de linha de AWS comando.

- Para obter detalhes da API, consulte DescribeAutoScalingInstancesem Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
   public static void describeAutoScalingInstance(AutoScalingClient
autoScalingClient, String id) {
       try {
           DescribeAutoScalingInstancesRequest
describeAutoScalingInstancesRequest = DescribeAutoScalingInstancesRequest
                   .builder()
                   .instanceIds(id)
                   .build();

           DescribeAutoScalingInstancesResponse response = autoScalingClient

.describeAutoScalingInstances(describeAutoScalingInstancesRequest);
           List<AutoScalingInstanceDetails> instances =
response.autoScalingInstances();
           for (AutoScalingInstanceDetails instance : instances) {
               System.out.println("The instance lifecycle state is: " +
instance.lifecycleState());
           }

       } catch (AutoScalingException e) {
           System.err.println(e.awsErrorDetails().errorMessage());
           System.exit(1);
       }
   }
```

- Para obter detalhes da API, consulte DescribeAutoScalingInstancesa Referência AWS SDK for Java 2.x da API.

Kotlin

SDK para Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun describeAutoScalingInstance(id: String) {
    val describeAutoScalingInstancesRequest =
        DescribeAutoScalingInstancesRequest {
            instanceIds = listOf(id)
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        val response =
 autoScalingClient.describeAutoScalingInstances(describeAutoScalingInstancesRequest)
        response.autoScalingInstances?.forEach { group ->
            println("The instance lifecycle state is: ${group.lifecycleState}")
        }
    }
}
```

- Para obter detalhes da API, consulte a DescribeAutoScalingInstancesreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function describeAutoScalingInstances($instanceIds)
```

```
    {
        return $this->autoScalingClient->describeAutoScalingInstances([
            'InstanceIds' => $instanceIds
        ]);
    }
```

- Para obter detalhes da API, consulte DescribeAutoScalingInstancesa Referência AWS SDK para PHP da API.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo lista as instâncias IDs do Auto Scaling.

```
Get-ASAutoScalingInstance | format-table -property InstanceId
```

Saída:

```
InstanceId
----------
i-12345678
i-87654321
i-abcd1234
```

Exemplo 2: Este exemplo descreve a instância do Auto Scaling especificada.

```
Get-ASAutoScalingInstance -InstanceId i-12345678
```

Saída:

```
AutoScalingGroupName    : my-asg
AvailabilityZone        : us-west-2b
HealthStatus            : HEALTHY
InstanceId              : i-12345678
LaunchConfigurationName : my-lc
LifecycleState          : InService
```

Exemplo 3: Este exemplo descreve duas instâncias do Auto Scaling especificadas.

```
Get-ASAutoScalingInstance -InstanceId @("i-12345678", "i-87654321")
```

Exemplo 4: Este exemplo descreve as instâncias do Auto Scaling do grupo do Auto Scaling especificado.

```
(Get-ASAutoScalingGroup -AutoScalingGroupName my-asg).Instances | Get-
ASAutoScalingInstance
```

Exemplo 5: Este exemplo descreve todas as instâncias do Auto Scaling.

```
Get-ASAutoScalingInstance
```

- Para obter detalhes da API, consulte DescribeAutoScalingInstancesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo lista as instâncias IDs do Auto Scaling.

```
Get-ASAutoScalingInstance | format-table -property InstanceId
```

Saída:

```
InstanceId
----------
i-12345678
i-87654321
i-abcd1234
```

Exemplo 2: Este exemplo descreve a instância do Auto Scaling especificada.

```
Get-ASAutoScalingInstance -InstanceId i-12345678
```

Saída:

```
AutoScalingGroupName    : my-asg
AvailabilityZone        : us-west-2b
HealthStatus            : HEALTHY
```

```
InstanceId              : i-12345678
LaunchConfigurationName : my-lc
LifecycleState          : InService
```

Exemplo 3: Este exemplo descreve duas instâncias do Auto Scaling especificadas.

```
Get-ASAutoScalingInstance -InstanceId @("i-12345678", "i-87654321")
```

Exemplo 4: Este exemplo descreve as instâncias do Auto Scaling do grupo do Auto Scaling especificado.

```
(Get-ASAutoScalingGroup -AutoScalingGroupName my-asg).Instances | Get-
ASAutoScalingInstance
```

Exemplo 5: Este exemplo descreve todas as instâncias do Auto Scaling.

```
Get-ASAutoScalingInstance
```

- Para obter detalhes da API, consulte DescribeAutoScalingInstancesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
```

```
        self.autoscaling_client = autoscaling_client


    def describe_instances(self, instance_ids: List[str]) -> List[Dict[str,
Any]]:
        """
        Gets information about instances.

        :param instance_ids: A list of instance IDs to look up.
        :return: A list of dictionaries with information about each instance,
                 or an empty list if none are found.
        :raises ClientError: If there is an error describing the instances.
        """
        try:
            paginator = self.autoscaling_client.get_paginator(
                "describe_auto_scaling_instances"
            )
            response_iterator = paginator.paginate(InstanceIds=instance_ids)

            instances = []
            for response in response_iterator:
                instances.extend(response.get("AutoScalingInstances", []))

            logger.info(f"Successfully described instances: {instance_ids}")

        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            logger.error(
                f"Couldn't describe instances {instance_ids}. Error code:
{error_code}, Message: {err.response['Error']['Message']}"
            )
            raise
        else:
            return instances
```

- Para obter detalhes da API, consulte a DescribeAutoScalingInstancesReferência da API AWS SDK for Python (Boto3).

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    pub async fn list_instances(&self) -> Result<Vec<String>, ScenarioError> {
        // The direct way to list instances is by using
 DescribeAutoScalingGroup's instances property. However, this returns a
 Vec<Instance>, as opposed to a Vec<AutoScalingInstanceDetails>.
        // Ok(self.get_group().await?.instances.unwrap_or_default().map(|
i| i.instance_id.clone().unwrap_or_default()).filter(|id| !
id.is_empty()).collect())

        // Alternatively, and for the sake of example,
 DescribeAutoScalingInstances returns a list that can be filtered by the client.
        self.autoscaling
            .describe_auto_scaling_instances()
            .into_paginator()
            .items()
            .send()
            .try_collect()
            .await
            .map(|items| {
                items
                    .into_iter()
                    .filter(|i| {
                        i.auto_scaling_group_name.as_deref()
                            == Some(self.auto_scaling_group_name.as_str())
                    })
                    .map(|i| i.instance_id.unwrap_or_default())
                    .filter(|id| !id.is_empty())
                    .collect::<Vec<String>>()
            })
            .map_err(|err| ScenarioError::new("Failed to get list of auto scaling
 instances", &err))
    }
```

- Para obter detalhes da API, consulte a DescribeAutoScalingInstancesreferência da API AWS SDK for Rust.

SAP ABAP

SDK para SAP ABAP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    " Example: it_instance_ids contains a list of instance IDs

    TRY.
        DATA(lo_output) = ao_asc->describeautoscalinginstances(
          it_instanceids = it_instance_ids ).

        ot_output = lo_output->get_autoscalinginstances( ).

        MESSAGE 'Auto Scaling instances information retrieved successfully' TYPE
'I'.

      CATCH /aws1/cx_ascresrccontionfault INTO DATA(lo_contention).
        RAISE EXCEPTION lo_contention.
      CATCH /aws1/cx_rt_generic INTO DATA(lo_generic_exception).
        RAISE EXCEPTION lo_generic_exception.
    ENDTRY.
```

- Para obter detalhes da API, consulte a DescribeAutoScalingInstancesreferência da API AWS SDK for SAP ABAP.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribeAutoScalingNotificationTypes`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeAutoScalingNotificationTypes`.

CLI

AWS CLI

Como descrever os tipos de notificação disponíveis

Este exemplo descreve os tipos de notificação disponíveis.

```
aws autoscaling describe-auto-scaling-notification-types
```

Saída:

```
{
    "AutoScalingNotificationTypes": [
        "autoscaling:EC2_INSTANCE_LAUNCH",
        "autoscaling:EC2_INSTANCE_LAUNCH_ERROR",
        "autoscaling:EC2_INSTANCE_TERMINATE",
        "autoscaling:EC2_INSTANCE_TERMINATE_ERROR",
        "autoscaling:TEST_NOTIFICATION"
    ]
}
```

Para obter mais informações, consulte Obter notificações do Amazon SNS quando o grupo de Auto Scaling é dimensionado no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DescribeAutoScalingNotificationTypesem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo lista os tipos de notificação compatíveis com o Auto Scaling.

```
Get-ASAutoScalingNotificationType
```

Saída:

```
autoscaling:EC2_INSTANCE_LAUNCH
autoscaling:EC2_INSTANCE_LAUNCH_ERROR
autoscaling:EC2_INSTANCE_TERMINATE
autoscaling:EC2_INSTANCE_TERMINATE_ERROR
autoscaling:TEST_NOTIFICATION
```

- Para obter detalhes da API, consulte DescribeAutoScalingNotificationTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo lista os tipos de notificação compatíveis com o Auto Scaling.

```
Get-ASAutoScalingNotificationType
```

Saída:

```
autoscaling:EC2_INSTANCE_LAUNCH
autoscaling:EC2_INSTANCE_LAUNCH_ERROR
autoscaling:EC2_INSTANCE_TERMINATE
autoscaling:EC2_INSTANCE_TERMINATE_ERROR
autoscaling:TEST_NOTIFICATION
```

- Para obter detalhes da API, consulte DescribeAutoScalingNotificationTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribeLaunchConfigurations`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeLaunchConfigurations`.

CLI

AWS CLI

Exemplo 1: Como descrever a configuração de execução especificada.

Este exemplo descreve a configuração de execução especificada.

```
aws autoscaling describe-launch-configurations \
    --launch-configuration-names my-launch-config
```

Saída:

```
{
    "LaunchConfigurations": [
        {
            "LaunchConfigurationName": "my-launch-config",
            "LaunchConfigurationARN": "arn:aws:autoscaling:us-
west-2:123456789012:launchConfiguration:98d3b196-4cf9-4e88-8ca1-8547c24ced8b:launchConfig
my-launch-config",
            "ImageId": "ami-0528a5175983e7f28",
            "KeyName": "my-key-pair-uswest2",
            "SecurityGroups": [
                "sg-05eaec502fcdadc2e"
            ],
            "ClassicLinkVPCSecurityGroups": [],
            "UserData": "",
            "InstanceType": "t2.micro",
            "KernelId": "",
            "RamdiskId": "",
            "BlockDeviceMappings": [
                {
                    "DeviceName": "/dev/xvda",
                    "Ebs": {
                        "SnapshotId": "snap-06c1606ba5ca274b1",
                        "VolumeSize": 8,
                        "VolumeType": "gp2",
                        "DeleteOnTermination": true,
                        "Encrypted": false
                    }
                }
            ],
            "InstanceMonitoring": {
                "Enabled": true
            },
            "CreatedTime": "2020-10-28T02:39:22.321Z",
            "EbsOptimized": false,
            "AssociatePublicIpAddress": true,
            "MetadataOptions": {
```

```
                "HttpTokens": "required",
                "HttpPutResponseHopLimit": 1,
                "HttpEndpoint": "disabled"
            }
        }
    ]
}
```

Exemplo 2: descrever um número especificado de configurações de execução

Para retornar um número específico de configurações de execução, use a opção `--max-items`.

```
aws autoscaling describe-launch-configurations \
    --max-items 1
```

Se a saída incluir um campo `NextToken`, há mais configurações de execução. Para obter as configurações de execução adicionais, use o valor desse campo com a opção `--starting-token` em uma chamada subsequente da seguinte maneira.

```
aws autoscaling describe-launch-configurations \
    --starting-token Z3M3LMPEXAMPLE
```

- Para obter detalhes da API, consulte DescribeLaunchConfigurationsem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo lista os nomes de suas configurações de execução.

```
Get-ASLaunchConfiguration | format-table -property LaunchConfigurationName
```

Saída:

```
LaunchConfigurationName
-----------------------
my-lc-1
```

```
my-lc-2
my-lc-3
my-lc-4
my-lc-5
```

Exemplo 2: Este exemplo descreve a configuração de execução especificada.

```
Get-ASLaunchConfiguration -LaunchConfigurationName my-lc-1
```

Saída:

```
AssociatePublicIpAddress     : True
BlockDeviceMappings          : {/dev/xvda}
ClassicLinkVPCId             :
ClassicLinkVPCSecurityGroups : {}
CreatedTime                  : 12/12/2014 3:22:08 PM
EbsOptimized                 : False
IamInstanceProfile           :
ImageId                      : ami-043a5034
InstanceMonitoring           : Amazon.AutoScaling.Model.InstanceMonitoring
InstanceType                 : t2.micro
KernelId                     :
KeyName                      :
LaunchConfigurationARN       : arn:aws:autoscaling:us-
west-2:123456789012:launchConfiguration:7e5f31e4-693b-4604-9322-
                               e6f68d7fafad:launchConfigurationName/my-lc-1
LaunchConfigurationName      : my-lc-1
PlacementTenancy             :
RamdiskId                    :
SecurityGroups               : {sg-67ef0308}
SpotPrice                    :
UserData                     :
```

Exemplo 3: Este exemplo descreve as duas configurações de execução especificadas.

```
Get-ASLaunchConfiguration -LaunchConfigurationName @("my-lc-1", "my-lc-2")
```

Exemplo 4: Este exemplo descreve todas as suas configurações de execução.

```
Get-ASLaunchConfiguration
```

- Para obter detalhes da API, consulte DescribeLaunchConfigurationsem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo lista os nomes de suas configurações de execução.

```
Get-ASLaunchConfiguration | format-table -property LaunchConfigurationName
```

Saída:

```
LaunchConfigurationName
-----------------------
my-lc-1
my-lc-2
my-lc-3
my-lc-4
my-lc-5
```

Exemplo 2: Este exemplo descreve a configuração de execução especificada.

```
Get-ASLaunchConfiguration -LaunchConfigurationName my-lc-1
```

Saída:

```
AssociatePublicIpAddress     : True
BlockDeviceMappings          : {/dev/xvda}
ClassicLinkVPCId             :
ClassicLinkVPCSecurityGroups : {}
CreatedTime                  : 12/12/2014 3:22:08 PM
EbsOptimized                 : False
IamInstanceProfile           :
ImageId                      : ami-043a5034
InstanceMonitoring           : Amazon.AutoScaling.Model.InstanceMonitoring
InstanceType                 : t2.micro
KernelId                     :
KeyName                      :
LaunchConfigurationARN       : arn:aws:autoscaling:us-
west-2:123456789012:launchConfiguration:7e5f31e4-693b-4604-9322-
                               e6f68d7fafad:launchConfigurationName/my-lc-1
LaunchConfigurationName      : my-lc-1
```

```
PlacementTenancy             :
RamdiskId                    :
SecurityGroups               : {sg-67ef0308}
SpotPrice                    :
UserData                     :
```

Exemplo 3: Este exemplo descreve as duas configurações de execução especificadas.

```
Get-ASLaunchConfiguration -LaunchConfigurationName @("my-lc-1", "my-lc-2")
```

Exemplo 4: Este exemplo descreve todas as suas configurações de execução.

```
Get-ASLaunchConfiguration
```

- Para obter detalhes da API, consulte DescribeLaunchConfigurationsem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribeLifecycleHookTypes`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeLifecycleHookTypes`.

CLI

AWS CLI

Descreve os tipos disponíveis de ganchos do ciclo de vida.

Este exemplo descreve os tipos de ganchos do ciclo de vida disponíveis.

```
aws autoscaling describe-lifecycle-hook-types
```

Saída:

```
{
    "LifecycleHookTypes": [
```

```
        "autoscaling:EC2_INSTANCE_LAUNCHING",
        "autoscaling:EC2_INSTANCE_TERMINATING"
    ]
}
```

- Para obter detalhes da API, consulte DescribeLifecycleHookTypesem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo lista os tipos de gancho do ciclo de vida compatíveis com o Auto Scaling.

```
Get-ASLifecycleHookType
```

Saída:

```
autoscaling:EC2_INSTANCE_LAUNCHING
auto-scaling:EC2_INSTANCE_TERMINATING
```

- Para obter detalhes da API, consulte DescribeLifecycleHookTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo lista os tipos de gancho do ciclo de vida compatíveis com o Auto Scaling.

```
Get-ASLifecycleHookType
```

Saída:

```
autoscaling:EC2_INSTANCE_LAUNCHING
auto-scaling:EC2_INSTANCE_TERMINATING
```

- Para obter detalhes da API, consulte DescribeLifecycleHookTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribeLifecycleHooks`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeLifecycleHooks`.

CLI

AWS CLI

Como descrever seus ganchos do ciclo de vida

Este exemplo descreve os ganchos do ciclo de vida do grupo do Auto Scaling especificado.

```
aws autoscaling describe-lifecycle-hooks \
    --auto-scaling-group-name my-asg
```

Saída:

```
{
    "LifecycleHooks": [
        {
            "GlobalTimeout": 3000,
            "HeartbeatTimeout": 30,
            "AutoScalingGroupName": "my-asg",
            "LifecycleHookName": "my-launch-hook",
            "DefaultResult": "ABANDON",
            "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING"
        },
        {
            "GlobalTimeout": 6000,
            "HeartbeatTimeout": 60,
            "AutoScalingGroupName": "my-asg",
            "LifecycleHookName": "my-termination-hook",
            "DefaultResult": "CONTINUE",
            "LifecycleTransition": "autoscaling:EC2_INSTANCE_TERMINATING"
        }
    ]
}
```

- Para obter detalhes da API, consulte DescribeLifecycleHooksem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo descreve o gancho do ciclo de vida especificado.

```
Get-ASLifecycleHook -AutoScalingGroupName my-asg -LifecycleHookName
 myLifecycleHook
```

Saída:

```
AutoScalingGroupName  : my-asg
DefaultResult         : ABANDON
GlobalTimeout         : 172800
HeartbeatTimeout      : 3600
LifecycleHookName     : myLifecycleHook
LifecycleTransition   : auto-scaling:EC2_INSTANCE_LAUNCHING
NotificationMetadata  :
NotificationTargetARN : arn:aws:sns:us-west-2:123456789012:my-topic
RoleARN               : arn:aws:iam::123456789012:role/my-iam-role
```

Exemplo 2: Descreve os ganchos do ciclo de vida do grupo de Auto Scaling especificado.

```
Get-ASLifecycleHook -AutoScalingGroupName my-asg
```

Exemplo 3: Este exemplo descreve todos ganchos do ciclo de vida os todos seus grupos do Auto Scaling.

```
Get-ASLifecycleHook
```

- Para obter detalhes da API, consulte DescribeLifecycleHooksem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo descreve o gancho do ciclo de vida especificado.

```
Get-ASLifecycleHook -AutoScalingGroupName my-asg -LifecycleHookName
 myLifecycleHook
```

Saída:

```
AutoScalingGroupName  : my-asg
DefaultResult         : ABANDON
GlobalTimeout         : 172800
HeartbeatTimeout      : 3600
LifecycleHookName     : myLifecycleHook
LifecycleTransition   : auto-scaling:EC2_INSTANCE_LAUNCHING
NotificationMetadata  :
NotificationTargetARN : arn:aws:sns:us-west-2:123456789012:my-topic
RoleARN               : arn:aws:iam::123456789012:role/my-iam-role
```

Exemplo 2: Descreve os ganchos do ciclo de vida do grupo de Auto Scaling especificado.

```
Get-ASLifecycleHook -AutoScalingGroupName my-asg
```

Exemplo 3: Este exemplo descreve todos ganchos do ciclo de vida os todos seus grupos do Auto Scaling.

```
Get-ASLifecycleHook
```

- Para obter detalhes da API, consulte DescribeLifecycleHooksem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribeLoadBalancers`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeLoadBalancers`.

CLI

AWS CLI

Como descrever os Classic Load Balancers de um grupo do Auto Scaling

Esse exemplo descreve os load balancers clássicos do grupo de Auto Scaling especificado.

```
aws autoscaling describe-load-balancers \
    --auto-scaling-group-name my-asg
```

Saída:

```
{
    "LoadBalancers": [
        {
            "State": "Added",
            "LoadBalancerName": "my-load-balancer"
        }
    ]
}
```

- Para obter detalhes da API, consulte DescribeLoadBalancersem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Descreve os load balancers do grupo de Auto Scaling especificado.

```
Get-ASLoadBalancer -AutoScalingGroupName my-asg
```

Saída:

```
LoadBalancerName    State
----------------    -----
my-lb               Added
```

- Para obter detalhes da API, consulte DescribeLoadBalancersem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Descreve os load balancers do grupo de Auto Scaling especificado.

```
Get-ASLoadBalancer -AutoScalingGroupName my-asg
```

Saída:

```
LoadBalancerName    State
----------------    -----
my-lb               Added
```

- Para obter detalhes da API, consulte DescribeLoadBalancersem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribeMetricCollectionTypes`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeMetricCollectionTypes`.

CLI

AWS CLI

Como descrever os tipos de coleta de métricas disponíveis

Este exemplo descreve os tipos de coleta de métricas disponíveis.

```
aws autoscaling describe-metric-collection-types
```

Saída:

```
{
    "Metrics": [
        {
            "Metric": "GroupMinSize"
        },
        {
            "Metric": "GroupMaxSize"
        },
        {
            "Metric": "GroupDesiredCapacity"
        },
```

```
        {
            "Metric": "GroupInServiceInstances"
        },
        {
            "Metric": "GroupInServiceCapacity"
        },
        {
            "Metric": "GroupPendingInstances"
        },
        {
            "Metric": "GroupPendingCapacity"
        },
        {
            "Metric": "GroupTerminatingInstances"
        },
        {
            "Metric": "GroupTerminatingCapacity"
        },
        {
            "Metric": "GroupStandbyInstances"
        },
        {
            "Metric": "GroupStandbyCapacity"
        },
        {
            "Metric": "GroupTotalInstances"
        },
        {
            "Metric": "GroupTotalCapacity"
        }
    ],
    "Granularities": [
        {
            "Granularity": "1Minute"
        }
    ]
}
```

Para obter mais informações, consulte Métricas de grupo do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DescribeMetricCollectionTypesem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo lista os tipos de coleta de métricas compatíveis com o Auto Scaling.

```
(Get-ASMetricCollectionType).Metrics
```

Saída:

```
Metric
------
GroupMinSize
GroupMaxSize
GroupDesiredCapacity
GroupInServiceInstances
GroupPendingInstances
GroupTerminatingInstances
GroupStandbyInstances
GroupTotalInstances
```

Exemplo 2: Este exemplo lista as granularidades correspondentes.

```
(Get-ASMetricCollectionType).Granularities
```

Saída:

```
Granularity
-----------
1Minute
```

- Para obter detalhes da API, consulte DescribeMetricCollectionTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo lista os tipos de coleta de métricas compatíveis com o Auto Scaling.

```
(Get-ASMetricCollectionType).Metrics
```

Saída:

```
Metric
------
GroupMinSize
GroupMaxSize
GroupDesiredCapacity
GroupInServiceInstances
GroupPendingInstances
GroupTerminatingInstances
GroupStandbyInstances
GroupTotalInstances
```

Exemplo 2: Este exemplo lista as granularidades correspondentes.

```
(Get-ASMetricCollectionType).Granularities
```

Saída:

```
Granularity
-----------
1Minute
```

- Para obter detalhes da API, consulte DescribeMetricCollectionTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribeNotificationConfigurations`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeNotificationConfigurations`.

CLI

AWS CLI

Exemplo 1: Como descrever as configurações de notificação de um grupo especificado.

Este exemplo descreve as configurações de notificação do grupo do Auto Scaling especificado.

```
aws autoscaling describe-notification-configurations \
    --auto-scaling-group-name my-asg
```

Saída:

```
{
    "NotificationConfigurations": [
        {
            "AutoScalingGroupName": "my-asg",
            "NotificationType": "autoscaling:TEST_NOTIFICATION",
            "TopicARN": "arn:aws:sns:us-west-2:123456789012:my-sns-topic-2"
        },
        {
            "AutoScalingGroupName": "my-asg",
            "NotificationType": "autoscaling:TEST_NOTIFICATION",
            "TopicARN": "arn:aws:sns:us-west-2:123456789012:my-sns-topic"
        }
    ]
}
```

Para obter mais informações, consulte Obter notificações do Amazon SNS quando o grupo de Auto Scaling é dimensionado no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 1: Como descrever um número especificado de configurações de notificação.

Para retornar um número especificado de configurações de notificação, use o parâmetro `max-items`.

```
aws autoscaling describe-notification-configurations \
    --auto-scaling-group-name my-auto-scaling-group \
    --max-items 1
```

Saída:

```
{
    "NotificationConfigurations": [
        {
            "AutoScalingGroupName": "my-asg",
            "NotificationType": "autoscaling:TEST_NOTIFICATION",
            "TopicARN": "arn:aws:sns:us-west-2:123456789012:my-sns-topic-2"
        },
```

```
        {
            "AutoScalingGroupName": "my-asg",
            "NotificationType": "autoscaling:TEST_NOTIFICATION",
            "TopicARN": "arn:aws:sns:us-west-2:123456789012:my-sns-topic"
        }
    ]
}
```

Se a saída incluir um campo `NextToken`, há mais configurações de notificação. Para obter as notificações adicionais, use o valor desse campo com o parâmetro `starting-token` em uma chamada subsequente da seguinte maneira.

```
aws autoscaling describe-notification-configurations \
    --auto-scaling-group-name my-asg \
    --starting-token Z3M3LMPEXAMPLE
```

Para obter mais informações, consulte Obter notificações do Amazon SNS quando o grupo de Auto Scaling é dimensionado no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DescribeNotificationConfigurationsem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Esse exemplo descreve as ações de notificação associadas ao grupo de Auto Scaling especificado.

```
Get-ASNotificationConfiguration -AutoScalingGroupName my-asg | format-list
```

Saída:

```
AutoScalingGroupName : my-asg
NotificationType     : auto-scaling:EC2_INSTANCE_LAUNCH
TopicARN             : arn:aws:sns:us-west-2:123456789012:my-topic

AutoScalingGroupName : my-asg
NotificationType     : auto-scaling:EC2_INSTANCE_TERMINATE
TopicARN             : arn:aws:sns:us-west-2:123456789012:my-topic
```

Exemplo 2: Este exemplo descreve as ações de notificação associadas a todos os seus grupos do Auto Scaling.

```
Get-ASNotificationConfiguration
```

- Para obter detalhes da API, consulte DescribeNotificationConfigurationsem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Esse exemplo descreve as ações de notificação associadas ao grupo de Auto Scaling especificado.

```
Get-ASNotificationConfiguration -AutoScalingGroupName my-asg | format-list
```

Saída:

```
AutoScalingGroupName : my-asg
NotificationType     : auto-scaling:EC2_INSTANCE_LAUNCH
TopicARN             : arn:aws:sns:us-west-2:123456789012:my-topic

AutoScalingGroupName : my-asg
NotificationType     : auto-scaling:EC2_INSTANCE_TERMINATE
TopicARN             : arn:aws:sns:us-west-2:123456789012:my-topic
```

Exemplo 2: Este exemplo descreve as ações de notificação associadas a todos os seus grupos do Auto Scaling.

```
Get-ASNotificationConfiguration
```

- Para obter detalhes da API, consulte DescribeNotificationConfigurationsem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribePolicies`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribePolicies`.

CLI

AWS CLI

Exemplo 1: como descrever as escalas do Auto Scaling especificado

Este exemplo descreve as politicas de escalabilidade do grupo do Auto Scaling especificado.

```
aws autoscaling describe-policies \
    --auto-scaling-group-name my-asg
```

Saída:

```
{
    "ScalingPolicies": [
        {
            "AutoScalingGroupName": "my-asg",
            "PolicyName": "alb1000-target-tracking-scaling-policy",
            "PolicyARN": "arn:aws:autoscaling:us-
west-2:123456789012:scalingPolicy:3065d9c8-9969-4bec-
bb6a-3fbe5550fde6:autoScalingGroupName/my-asg:policyName/alb1000-target-tracking-
scaling-policy",
            "PolicyType": "TargetTrackingScaling",
            "StepAdjustments": [],
            "Alarms": [
                {
                    "AlarmName": "TargetTracking-my-asg-
AlarmHigh-924887a9-12d7-4e01-8686-6f844d13a196",
                    "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-
AlarmHigh-924887a9-12d7-4e01-8686-6f844d13a196"
                },
                {
                    "AlarmName": "TargetTracking-my-asg-AlarmLow-f96f899d-
b8e7-4d09-a010-c1aaa35da296",
                    "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-f96f899d-b8e7-4d09-a010-
c1aaa35da296"
                }
            ],
            "TargetTrackingConfiguration": {
                "PredefinedMetricSpecification": {
                    "PredefinedMetricType": "ALBRequestCountPerTarget",
```

```
                    "ResourceLabel": "app/my-alb/778d41231b141a0f/targetgroup/my-
alb-target-group/943f017f100becff"
                },
                "TargetValue": 1000.0,
                "DisableScaleIn": false
            },
            "Enabled": true
        },
        {
            "AutoScalingGroupName": "my-asg",
            "PolicyName": "cpu40-target-tracking-scaling-policy",
            "PolicyARN": "arn:aws:autoscaling:us-
west-2:123456789012:scalingPolicy:5fd26f71-39d4-4690-82a9-
b8515c45cdde:autoScalingGroupName/my-asg:policyName/cpu40-target-tracking-
scaling-policy",
            "PolicyType": "TargetTrackingScaling",
            "StepAdjustments": [],
            "Alarms": [
                {
                    "AlarmName": "TargetTracking-my-asg-
AlarmHigh-139f9789-37b9-42ad-bea5-b5b147d7f473",
                    "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-139f9789-37b9-42ad-
bea5-b5b147d7f473"
                },
                {
                    "AlarmName": "TargetTracking-my-asg-AlarmLow-bd681c67-
fc18-4c56-8468-fb8e413009c9",
                    "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-bd681c67-fc18-4c56-8468-
fb8e413009c9"
                }
            ],
            "TargetTrackingConfiguration": {
                "PredefinedMetricSpecification": {
                    "PredefinedMetricType": "ASGAverageCPUUtilization"
                },
                "TargetValue": 40.0,
                "DisableScaleIn": false
            },
            "Enabled": true
        }
    ]
```

```
}
```

Para obter mais informações, consulte Escalabilidade Dinâmica no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: Como descrever as políticas de escalabilidade de um nome especificado.

Para retornar políticas de escalabilidade específicas, use a opção `--policy-names`.

```
aws autoscaling describe-policies \
    --auto-scaling-group-name my-asg \
    --policy-names cpu40-target-tracking-scaling-policy
```

Consulte uma amostra de saída no exemplo 1.

Para obter mais informações, consulte Escalabilidade dinâmica no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 3: Como descrever várias políticas de escalabilidade.

Use a opção `--max-items` para retornar um número específico de politicas.

```
aws autoscaling describe-policies \
    --auto-scaling-group-name my-asg \
    --max-items 1
```

Consulte uma amostra de saída no exemplo 1.

Se a saída incluir um campo `NextToken`, use o valor desse campo com a opção `--starting-token` em uma chamada subsequente para obter as políticas adicionais.

```
aws autoscaling describe-policies --auto-scaling-group-name my-asg --starting-
token Z3M3LMPEXAMPLE
```

Para obter mais informações, consulte Escalabilidade dinâmica no Manual do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DescribePoliciesem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Esse exemplo descreve o grupo do Auto Scaling especificado.

```
Get-ASPolicy -AutoScalingGroupName my-asg
```

Saída:

```
AdjustmentType          : ChangeInCapacity
Alarms                  : {}
AutoScalingGroupName    : my-asg
Cooldown                : 0
EstimatedInstanceWarmup : 0
MetricAggregationType   :
MinAdjustmentMagnitude  : 0
MinAdjustmentStep       : 0
PolicyARN               : arn:aws:auto-scaling:us-
west-2:123456789012:scalingPolicy:aa3836ab-5462-42c7-adab-e1d769fc24ef
                          :autoScalingGroupName/my-asg:policyName/myScaleInPolicy
PolicyName              : myScaleInPolicy
PolicyType              : SimpleScaling
ScalingAdjustment       : -1
StepAdjustments         : {}
```

Exemplo 2: Este exemplo descreve duas políticas especificadas para o grupo do Auto Scaling especificado.

```
Get-ASPolicy -AutoScalingGroupName my-asg -PolicyName @("myScaleOutPolicy",
 "myScaleInPolicy")
```

Exemplo 3: Este exemplo descreve todas as políticas para todos os seus grupos do Auto Scaling.

```
Get-ASPolicy
```

- Para obter detalhes da API, consulte DescribePoliciesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Esse exemplo descreve o grupo do Auto Scaling especificado.

```
Get-ASPolicy -AutoScalingGroupName my-asg
```

Saída:

```
AdjustmentType          : ChangeInCapacity
Alarms                  : {}
AutoScalingGroupName    : my-asg
Cooldown                : 0
EstimatedInstanceWarmup : 0
MetricAggregationType   :
MinAdjustmentMagnitude  : 0
MinAdjustmentStep       : 0
PolicyARN               : arn:aws:auto-scaling:us-
west-2:123456789012:scalingPolicy:aa3836ab-5462-42c7-adab-e1d769fc24ef
                          :autoScalingGroupName/my-asg:policyName/myScaleInPolicy
PolicyName              : myScaleInPolicy
PolicyType              : SimpleScaling
ScalingAdjustment       : -1
StepAdjustments         : {}
```

Exemplo 2: Este exemplo descreve duas políticas especificadas para o grupo do Auto Scaling especificado.

```
Get-ASPolicy -AutoScalingGroupName my-asg -PolicyName @("myScaleOutPolicy",
 "myScaleInPolicy")
```

Exemplo 3: Este exemplo descreve todas as políticas para todos os seus grupos do Auto Scaling.

```
Get-ASPolicy
```

- Para obter detalhes da API, consulte DescribePoliciesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Use **`DescribeScalingActivities`** com um AWS SDK ou CLI

Os exemplos de código a seguir mostram como usar o `DescribeScalingActivities`.

Exemplos de ações são trechos de código de programas maiores e devem ser executados em contexto. É possível ver essa ação em contexto no seguinte exemplo de código:

- Conheça os conceitos básicos

.NET

SDK para .NET (v4)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    /// <summary>
    /// Retrieve a list of the Amazon EC2 Auto Scaling activities for an
    /// Amazon EC2 Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A list of Amazon EC2 Auto Scaling activities.</returns>
    public async Task<List<Activity>> DescribeScalingActivitiesAsync(
        string groupName)
    {
        var activities = new List<Activity>();
        var scalingActivitiesRequest = new DescribeScalingActivitiesRequest
        {
            AutoScalingGroupName = groupName,
            MaxRecords = 10,
        };
```

```
        var response = await
_amazonAutoScaling.DescribeScalingActivitiesAsync(scalingActivitiesRequest);
        if (response.Activities != null)
        {
            activities = response.Activities;
        }
        return activities;
    }
```

- Para obter detalhes da API, consulte DescribeScalingActivitiesa Referência AWS SDK para .NET da API.

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

        Aws::AutoScaling::Model::DescribeScalingActivitiesRequest request;
        request.SetAutoScalingGroupName(groupName);

        Aws::Vector<Aws::AutoScaling::Model::Activity> allActivities;
        Aws::String nextToken; // Used for pagination;
        do {
            if (!nextToken.empty()) {
                request.SetNextToken(nextToken);
            }
            Aws::AutoScaling::Model::DescribeScalingActivitiesOutcome outcome =
                    autoScalingClient.DescribeScalingActivities(request);
```

```
            if (outcome.IsSuccess()) {
                const Aws::Vector<Aws::AutoScaling::Model::Activity> &activities
=
                        outcome.GetResult().GetActivities();
                allActivities.insert(allActivities.end(), activities.begin(),
activities.end());
                nextToken  = outcome.GetResult().GetNextToken();
            }
            else {
                std::cerr << "Error with AutoScaling::DescribeScalingActivities.
"
                          << outcome.GetError().GetMessage()
                          << std::endl;

            }
        } while (!nextToken.empty());

        std::cout << "Found " << allActivities.size() << " activities."
                  << std::endl;
        std::cout << "Activities are ordered with the most recent first."
                  << std::endl;
        for (const Aws::AutoScaling::Model::Activity &activity: allActivities) {
            std::cout << activity.GetDescription() << std::endl;
            std::cout << activity.GetDetails() << std::endl;
        }
```

- Para obter detalhes da API, consulte DescribeScalingActivitiesa Referência AWS SDK para C++ da API.

CLI

AWS CLI

Exemplo 1: como descrever as ações de escalabilidade do grupo especificado

Este exemplo descreve as ações de escalabilidade do grupo do Auto Scaling especificado.

```
aws autoscaling describe-scaling-activities \
    --auto-scaling-group-name my-asg
```

Saída:

```
{
    "Activities": [
        {
            "ActivityId": "f9f2d65b-f1f2-43e7-b46d-d86756459699",
            "Description": "Launching a new EC2 instance: i-0d44425630326060f",
            "AutoScalingGroupName": "my-asg",
            "Cause": "At 2020-10-30T19:35:51Z a user request update of
 AutoScalingGroup constraints to min: 0, max: 16, desired: 16 changing the
 desired capacity from 0 to 16.  At 2020-10-30T19:36:07Z an instance was started
 in response to a difference between desired and actual capacity, increasing the
 capacity from 0 to 16.",
            "StartTime": "2020-10-30T19:36:09.766Z",
            "EndTime": "2020-10-30T19:36:41Z",
            "StatusCode": "Successful",
            "Progress": 100,
            "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":
\"us-west-2b\"}"
        }
    ]
}
```

Para mais informações, consulte Verificar uma ação de escalabilidade para um grupo do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: como descrever as ações de escalabilidade de um grupo excluído

Adicione a opção `--include-deleted-groups` para descrever as ações de escalabilidade depois que o grupo do Auto Scaling for excluído.

```
aws autoscaling describe-scaling-activities \
    --auto-scaling-group-name my-asg \
    --include-deleted-groups
```

Saída:

```
{
    "Activities": [
        {
            "ActivityId": "e1f5de0e-f93e-1417-34ac-092a76fba220",
```

```
            "Description": "Launching a new EC2 instance.  Status Reason: Your
 Spot request price of 0.001 is lower than the minimum required Spot request
 fulfillment price of 0.0031. Launching EC2 instance failed.",
            "AutoScalingGroupName": "my-asg",
            "Cause": "At 2021-01-13T20:47:24Z a user request update of
 AutoScalingGroup constraints to min: 1, max: 5, desired: 3 changing the desired
 capacity from 0 to 3.  At 2021-01-13T20:47:27Z an instance was started in
 response to a difference between desired and actual capacity, increasing the
 capacity from 0 to 3.",
            "StartTime": "2021-01-13T20:47:30.094Z",
            "EndTime": "2021-01-13T20:47:30Z",
            "StatusCode": "Failed",
            "StatusMessage": "Your Spot request price of 0.001 is lower than
 the minimum required Spot request fulfillment price of 0.0031. Launching EC2
 instance failed.",
            "Progress": 100,
            "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":
\"us-west-2b\"}",
            "AutoScalingGroupState": "Deleted",
            "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
        }
    ]
}
```

Para obter mais informações, consulte Solucionar problemas do Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 3: como descrever um número especificado de ações de escalabilidade

Use a opção `--max-items` para retornar um número específico de ações.

```
aws autoscaling describe-scaling-activities \
    --max-items 1
```

Saída:

```
{
    "Activities": [
        {
            "ActivityId": "f9f2d65b-f1f2-43e7-b46d-d86756459699",
            "Description": "Launching a new EC2 instance: i-0d44425630326060f",
```

```
            "AutoScalingGroupName": "my-asg",
            "Cause": "At 2020-10-30T19:35:51Z a user request update of
 AutoScalingGroup constraints to min: 0, max: 16, desired: 16 changing the
 desired capacity from 0 to 16.  At 2020-10-30T19:36:07Z an instance was started
 in response to a difference between desired and actual capacity, increasing the
 capacity from 0 to 16.",
            "StartTime": "2020-10-30T19:36:09.766Z",
            "EndTime": "2020-10-30T19:36:41Z",
            "StatusCode": "Successful",
            "Progress": 100,
            "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":
\"us-west-2b\"}"
        }
    ]
}
```

Se a saída incluir um campo `NextToken`, há mais ações. Para obter as ações adicionais, use o valor desse campo com a opção `--starting-token` em uma chamada subsequente da seguinte maneira.

```
aws autoscaling describe-scaling-activities \
    --starting-token Z3M3LMPEXAMPLE
```

Para mais informações, consulte Verificar uma ação de escalabilidade para um grupo do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DescribeScalingActivitiesem Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public static void describeScalingActivities(AutoScalingClient
 autoScalingClient, String groupName) {
```

```
        try {
            DescribeScalingActivitiesRequest scalingActivitiesRequest =
 DescribeScalingActivitiesRequest.builder()
                    .autoScalingGroupName(groupName)
                    .maxRecords(10)
                    .build();

            DescribeScalingActivitiesResponse response = autoScalingClient
                    .describeScalingActivities(scalingActivitiesRequest);
            List<Activity> activities = response.activities();
            for (Activity activity : activities) {
                System.out.println("The activity Id is " +
 activity.activityId());
                System.out.println("The activity details are " +
 activity.details());
            }

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
```

- Para obter detalhes da API, consulte DescribeScalingActivitiesa Referência AWS SDK for Java 2.x da API.

Kotlin

SDK para Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun describeAutoScalingGroups(groupName: String) {
    val groupsReques =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
```

```
            maxRecords = 10
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        val response = autoScalingClient.describeAutoScalingGroups(groupsReques)
        response.autoScalingGroups?.forEach { group ->
            println("The service to use for the health checks:
 ${group.healthCheckType}")
        }
    }
}
```

- Para obter detalhes da API, consulte a DescribeScalingActivitiesreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function describeScalingActivities($autoScalingGroupName)
    {
        return $this->autoScalingClient->describeScalingActivities([
            'AutoScalingGroupName' => $autoScalingGroupName,
        ]);
    }
```

- Para obter detalhes da API, consulte DescribeScalingActivitiesa Referência AWS SDK para PHP da API.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo descreve as ações de escalabilidade das últimas seis semanas do grupo do Auto Scaling especificado.

```
Get-ASScalingActivity -AutoScalingGroupName my-asg
```

Saída:

```
ActivityId           : 063308ae-aa22-4a9b-94f4-9fae4EXAMPLE
AutoScalingGroupName : my-asg
Cause                : At 2015-11-22T15:45:16Z a user request explicitly set
 group desired capacity changing the desired
                       capacity from 1 to 2.  At 2015-11-22T15:45:34Z an instance
 was started in response to a difference
                       between desired and actual capacity, increasing the
 capacity from 1 to 2.
Description          : Launching a new EC2 instance: i-26e715fc
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              : 11/22/2015 7:46:09 AM
Progress             : 100
StartTime            : 11/22/2015 7:45:35 AM
StatusCode           : Successful
StatusMessage        :

ActivityId           : ce719997-086d-4c73-a2f1-ab703EXAMPLE
AutoScalingGroupName : my-asg
Cause                : At 2015-11-20T22:57:53Z a user request created an
 AutoScalingGroup changing the desired capacity
                        from 0 to 1.  At 2015-11-20T22:57:58Z an instance was
 started in response to a difference betwe
                       en desired and actual capacity, increasing the capacity
 from 0 to 1.
Description          : Launching a new EC2 instance: i-93633f9b
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              : 11/20/2015 2:58:32 PM
Progress             : 100
StartTime            : 11/20/2015 2:57:59 PM
StatusCode           : Successful
```

```
StatusMessage        :
```

Exemplo 2: Este exemplo descreve a atividade de escala especificada.

```
Get-ASScalingActivity -ActivityId "063308ae-aa22-4a9b-94f4-9fae4EXAMPLE"
```

Exemplo 3: Este exemplo descreve as atividades de escala das últimas seis semanas para todos os seus grupos do Auto Scaling.

```
Get-ASScalingActivity
```

- Para obter detalhes da API, consulte DescribeScalingActivitiesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo descreve as ações de escalabilidade das últimas seis semanas do grupo do Auto Scaling especificado.

```
Get-ASScalingActivity -AutoScalingGroupName my-asg
```

Saída:

```
ActivityId           : 063308ae-aa22-4a9b-94f4-9fae4EXAMPLE
AutoScalingGroupName : my-asg
Cause                : At 2015-11-22T15:45:16Z a user request explicitly set
 group desired capacity changing the desired
                       capacity from 1 to 2.  At 2015-11-22T15:45:34Z an instance
 was started in response to a difference
                       between desired and actual capacity, increasing the
 capacity from 1 to 2.
Description          : Launching a new EC2 instance: i-26e715fc
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              : 11/22/2015 7:46:09 AM
Progress             : 100
StartTime            : 11/22/2015 7:45:35 AM
StatusCode           : Successful
StatusMessage        :

ActivityId           : ce719997-086d-4c73-a2f1-ab703EXAMPLE
AutoScalingGroupName : my-asg
```

```
Cause                : At 2015-11-20T22:57:53Z a user request created an
 AutoScalingGroup changing the desired capacity
                       from 0 to 1.  At 2015-11-20T22:57:58Z an instance was
 started in response to a difference betwe
                      en desired and actual capacity, increasing the capacity
 from 0 to 1.
Description          : Launching a new EC2 instance: i-93633f9b
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              : 11/20/2015 2:58:32 PM
Progress             : 100
StartTime            : 11/20/2015 2:57:59 PM
StatusCode           : Successful
StatusMessage        :
```

Exemplo 2: Este exemplo descreve a atividade de escala especificada.

```
Get-ASScalingActivity -ActivityId "063308ae-aa22-4a9b-94f4-9fae4EXAMPLE"
```

Exemplo 3: Este exemplo descreve as atividades de escala das últimas seis semanas para todos os seus grupos do Auto Scaling.

```
Get-ASScalingActivity
```

- Para obter detalhes da API, consulte DescribeScalingActivitiesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""
```

```
    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def describe_scaling_activities(self, group_name: str) -> List[Dict[str,
Any]]:
        """
        Gets information about scaling activities for the group. Scaling
activities
        are things like instances stopping or starting in response to user
requests
        or capacity changes.

        :param group_name: The name of the group to look up.
        :return: A list of dictionaries representing the scaling activities for
the
                 group, ordered with the most recent activity first.
        :raises ClientError: If there is an error describing the scaling
activities.
        """
        try:
            paginator = self.autoscaling_client.get_paginator(
                "describe_scaling_activities"
            )
            response_iterator =
paginator.paginate(AutoScalingGroupName=group_name)
            activities = []
            for response in response_iterator:
                activities.extend(response.get("Activities", []))

            logger.info(
                f"Successfully described scaling activities for group
'{group_name}'."
            )

        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            logger.error(
                f"Couldn't describe scaling activities for group '{group_name}'.
Error code: {error_code}, Message: {err.response['Error']['Message']}"
            )
```

```
            if error_code == "ResourceContentionFault":
                logger.error(
                    f"There is a conflict with another operation that is
modifying the Auto Scaling group '{group_name}'. "
                    "Please try again later."
                )
            raise
        else:
            return activities
```

- Para obter detalhes da API, consulte a DescribeScalingActivitiesReferência da API AWS SDK for Python (Boto3).

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    pub async fn describe_scenario(&self) -> AutoScalingScenarioDescription {
        let group = self
            .autoscaling
            .describe_auto_scaling_groups()
            .auto_scaling_group_names(self.auto_scaling_group_name.clone())
            .send()
            .await
            .map(|s| {
                s.auto_scaling_groups()
                    .iter()
                    .map(|s| {
                        format!(
                            "{}: {}",
                            s.auto_scaling_group_name().unwrap_or("Unknown"),
                            s.status().unwrap_or("Unknown")
```

```
                        )
                })
                .collect::<Vec<String>>()
        })
        .map_err(|e| {
            ScenarioError::new("Failed to describe auto scaling groups for
scenario", &e)
        });

    let instances = self
        .list_instances()
        .await
        .map_err(|e| anyhow!("There was an error listing instances: {e}",));

    // 10. DescribeScalingActivities: list the scaling activities that have
occurred for the group so far.
    //   Bonus: use CloudWatch API to get and show some metrics collected for
the group.
    //   CW.ListMetrics with Namespace='AWS/AutoScaling' and
Dimensions=[{'Name': 'AutoScalingGroupName', 'Value': }]
    //   CW.GetMetricStatistics with Statistics='Sum'. Start and End times
must be in UTC!
    let activities = self
        .autoscaling
        .describe_scaling_activities()
        .auto_scaling_group_name(self.auto_scaling_group_name.clone())
        .into_paginator()
        .items()
        .send()
        .collect::<Result<Vec<_>, _>>()
        .await
        .map_err(|e| {
            anyhow!(
                "There was an error retrieving scaling activities: {}",
                DisplayErrorContext(&e)
            )
        });

    AutoScalingScenarioDescription {
        group,
        instances,
        activities,
    }
}
```

- Para obter detalhes da API, consulte a DescribeScalingActivitiesreferência da API AWS SDK for Rust.

SAP ABAP

SDK para SAP ABAP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
" Example: iv_group_name = 'my-auto-scaling-group'

TRY.
    DATA(lo_output) = ao_asc->describescalingactivities(
      iv_autoscalinggroupname = iv_group_name ).

    ot_output = lo_output->get_activities( ).

    MESSAGE 'Scaling activities retrieved successfully' TYPE 'I'.

  CATCH /aws1/cx_ascresrccontionfault INTO DATA(lo_contention).
    RAISE EXCEPTION lo_contention.
  CATCH /aws1/cx_rt_generic INTO DATA(lo_generic_exception).
    RAISE EXCEPTION lo_generic_exception.
ENDTRY.
```

- Para obter detalhes da API, consulte a DescribeScalingActivitiesreferência da API AWS SDK for SAP ABAP.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribeScalingProcessTypes`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeScalingProcessTypes`.

CLI

AWS CLI

Como descrever os tipos de processo disponíveis

Este exemplo descreve os tipos de processo disponíveis.

```
aws autoscaling describe-scaling-process-types
```

Saída:

```
{
    "Processes": [
        {
            "ProcessName": "AZRebalance"
        },
        {
            "ProcessName": "AddToLoadBalancer"
        },
        {
            "ProcessName": "AlarmNotification"
        },
        {
            "ProcessName": "HealthCheck"
        },
        {
            "ProcessName": "InstanceRefresh"
        },
        {
            "ProcessName": "Launch"
        },
        {
            "ProcessName": "ReplaceUnhealthy"
        },
        {
            "ProcessName": "ScheduledActions"
        },
        {
```

```
            "ProcessName": "Terminate"
        }
    ]
}
```

Para obter mais informações, consulte Suspensão e retomada dos processos de escalabilidade no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DescribeScalingProcessTypesem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo lista os tipos de processo compatíveis com o Auto Scaling.

```
Get-ASScalingProcessType
```

Saída:

```
ProcessName
-----------
AZRebalance
AddToLoadBalancer
AlarmNotification
HealthCheck
Launch
ReplaceUnhealthy
ScheduledActions
Terminate
```

- Para obter detalhes da API, consulte DescribeScalingProcessTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo lista os tipos de processo compatíveis com o Auto Scaling.

```
Get-ASScalingProcessType
```

Saída:

```
ProcessName
-----------
AZRebalance
AddToLoadBalancer
AlarmNotification
HealthCheck
Launch
ReplaceUnhealthy
ScheduledActions
Terminate
```

- Para obter detalhes da API, consulte DescribeScalingProcessTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribeScheduledActions`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeScheduledActions`.

CLI

AWS CLI

Exemplo 1: Como descrever todas as ações programadas.

Este exemplo descreve todas as ações programadas.

```
aws autoscaling describe-scheduled-actions
```

Saída:

```
{
    "ScheduledUpdateGroupActions": [
        {
            "AutoScalingGroupName": "my-asg",
            "ScheduledActionName": "my-recurring-action",
            "Recurrence": "30 0 1 1,6,12 *",
```

```
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledUpdateGroupAction:8e86b655-b2e6-4410-8f29-
b4f094d6871c:autoScalingGroupName/my-asg:scheduledActionName/my-recurring-
action",
            "StartTime": "2023-12-01T04:00:00Z",
            "Time": "2023-12-01T04:00:00Z",
            "MinSize": 1,
            "MaxSize": 6,
            "DesiredCapacity": 4,
            "TimeZone": "America/New_York"
        }
    ]
}
```

Para obter mais informações, consulte Escalabilidade programada no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: Como descrever as ações agendadas do grupo especificado

Para descrever as ações programadas para um grupo do Auto Scaling específico, use a opção `--auto-scaling-group-name`.

```
aws autoscaling describe-scheduled-actions \
    --auto-scaling-group-name my-asg
```

Saída:

```
{
    "ScheduledUpdateGroupActions": [
        {
            "AutoScalingGroupName": "my-asg",
            "ScheduledActionName": "my-recurring-action",
            "Recurrence": "30 0 1 1,6,12 *",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledUpdateGroupAction:8e86b655-b2e6-4410-8f29-
b4f094d6871c:autoScalingGroupName/my-asg:scheduledActionName/my-recurring-
action",
            "StartTime": "2023-12-01T04:00:00Z",
            "Time": "2023-12-01T04:00:00Z",
            "MinSize": 1,
            "MaxSize": 6,
            "DesiredCapacity": 4,
```

```
            "TimeZone": "America/New_York"
        }
    ]
}
```

Para obter mais informações, consulte Escalabilidade programada no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 3: Como descrever a ação programada especificada.

Para descrever uma ação programada específica, use a opção `--scheduled-action-names`.

```
aws autoscaling describe-scheduled-actions \
    --scheduled-action-names my-recurring-action
```

Saída:

```
{
    "ScheduledUpdateGroupActions": [
        {
            "AutoScalingGroupName": "my-asg",
            "ScheduledActionName": "my-recurring-action",
            "Recurrence": "30 0 1 1,6,12 *",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledUpdateGroupAction:8e86b655-b2e6-4410-8f29-
b4f094d6871c:autoScalingGroupName/my-asg:scheduledActionName/my-recurring-
action",
            "StartTime": "2023-12-01T04:00:00Z",
            "Time": "2023-12-01T04:00:00Z",
            "MinSize": 1,
            "MaxSize": 6,
            "DesiredCapacity": 4,
            "TimeZone": "America/New_York"
        }
    ]
}
```

Para obter mais informações, consulte Escalabilidade programada no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 4: Como descrever ações programadas com um horário de início especificado.

Para descrever as ações programadas que começam em um horário específico, use a opção `--start-time`.

```
aws autoscaling describe-scheduled-actions \
    --start-time "2023-12-01T04:00:00Z"
```

Saída:

```
{
    "ScheduledUpdateGroupActions": [
        {
            "AutoScalingGroupName": "my-asg",
            "ScheduledActionName": "my-recurring-action",
            "Recurrence": "30 0 1 1,6,12 *",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledUpdateGroupAction:8e86b655-b2e6-4410-8f29-
b4f094d6871c:autoScalingGroupName/my-asg:scheduledActionName/my-recurring-
action",
            "StartTime": "2023-12-01T04:00:00Z",
            "Time": "2023-12-01T04:00:00Z",
            "MinSize": 1,
            "MaxSize": 6,
            "DesiredCapacity": 4,
            "TimeZone": "America/New_York"
        }
    ]
}
```

Para obter mais informações, consulte Escalabilidade programada no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 5: Como descrever ações programadas que terminam em um horário especificado.

Para descrever as ações agendadas que terminam em um horário específico, use a opção `--end-time` .

```
aws autoscaling describe-scheduled-actions \
    --end-time "2023-12-01T04:00:00Z"
```

Saída:

```
{
```

```
    "ScheduledUpdateGroupActions": [
        {
            "AutoScalingGroupName": "my-asg",
            "ScheduledActionName": "my-recurring-action",
            "Recurrence": "30 0 1 1,6,12 *",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledUpdateGroupAction:8e86b655-b2e6-4410-8f29-
b4f094d6871c:autoScalingGroupName/my-asg:scheduledActionName/my-recurring-
action",
            "StartTime": "2023-12-01T04:00:00Z",
            "Time": "2023-12-01T04:00:00Z",
            "MinSize": 1,
            "MaxSize": 6,
            "DesiredCapacity": 4,
            "TimeZone": "America/New_York"
        }
    ]
}
```

Para obter mais informações, consulte Escalabilidade programada no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 6: como descrever um número especificado de ações agendadas

Use a opção `--max-items` para retornar um número específico de ações agendadas.

```
aws autoscaling describe-scheduled-actions \
    --auto-scaling-group-name my-asg \
    --max-items 1
```

Saída:

```
{
    "ScheduledUpdateGroupActions": [
        {
            "AutoScalingGroupName": "my-asg",
            "ScheduledActionName": "my-recurring-action",
            "Recurrence": "30 0 1 1,6,12 *",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledUpdateGroupAction:8e86b655-b2e6-4410-8f29-
b4f094d6871c:autoScalingGroupName/my-asg:scheduledActionName/my-recurring-
action",
            "StartTime": "2023-12-01T04:00:00Z",
```

```
            "Time": "2023-12-01T04:00:00Z",
            "MinSize": 1,
            "MaxSize": 6,
            "DesiredCapacity": 4,
            "TimeZone": "America/New_York"
        }
    ]
}
```

Se a saída incluir um campo `NextToken`, há mais ações agendadas. Para obter as ações agendadas, use o valor desse campo com a opção `--starting-token` em uma chamada subsequente da seguinte maneira.

```
aws autoscaling describe-scheduled-actions \
    --auto-scaling-group-name my-asg \
    --starting-token Z3M3LMPEXAMPLE
```

Para obter mais informações, consulte Escalabilidade programada no Manual do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DescribeScheduledActionsem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo descreve as ações de escalabilidade agendadas do grupo do Auto Scaling especificado.

```
Get-ASScheduledAction -AutoScalingGroupName my-asg
```

Saída:

```
AutoScalingGroupName : my-asg
DesiredCapacity      : 10
EndTime              :
MaxSize              :
MinSize              :
Recurrence           :
```

```
ScheduledActionARN   : arn:aws:autoscaling:us-
west-2:123456789012:scheduledUpdateGroupAction:8a4c5f24-6ec6-4306-a2dd-f7
                       2c3af3a4d6:autoScalingGroupName/my-
asg:scheduledActionName/myScheduledAction
ScheduledActionName  : myScheduledAction
StartTime            : 11/30/2015 8:00:00 AM
Time                 : 11/30/2015 8:00:00 AM
```

Exemplo 2: Este exemplo descreve as ações de escala programada especificadas.

```
Get-ASScheduledAction -ScheduledActionName @("myScheduledScaleOut",
 "myScheduledScaleIn")
```

Exemplo 3: Este exemplo descreve as ações de escala programadas que começam no horário especificado.

```
Get-ASScheduledAction -StartTime "2015-12-01T08:00:00Z"
```

Exemplo 4: Este exemplo descreve as ações de escala programadas que terminam no horário especificado.

```
Get-ASScheduledAction -EndTime "2015-12-30T08:00:00Z"
```

Exemplo 5: Este exemplo descreve as ações de escala programadas para todos os seus grupos do Auto Scaling.

```
Get-ASScheduledAction
```

- Para obter detalhes da API, consulte DescribeScheduledActionsem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo descreve as ações de escalabilidade agendadas do grupo do Auto Scaling especificado.

```
Get-ASScheduledAction -AutoScalingGroupName my-asg
```

Saída:

```
AutoScalingGroupName : my-asg
```

```
DesiredCapacity      : 10
EndTime              :
MaxSize              :
MinSize              :
Recurrence           :
ScheduledActionARN   : arn:aws:autoscaling:us-
west-2:123456789012:scheduledUpdateGroupAction:8a4c5f24-6ec6-4306-a2dd-f7
                       2c3af3a4d6:autoScalingGroupName/my-
asg:scheduledActionName/myScheduledAction
ScheduledActionName  : myScheduledAction
StartTime            : 11/30/2015 8:00:00 AM
Time                 : 11/30/2015 8:00:00 AM
```

Exemplo 2: Este exemplo descreve as ações de escala programada especificadas.

```
Get-ASScheduledAction -ScheduledActionName @("myScheduledScaleOut",
 "myScheduledScaleIn")
```

Exemplo 3: Este exemplo descreve as ações de escala programadas que começam no horário especificado.

```
Get-ASScheduledAction -StartTime "2015-12-01T08:00:00Z"
```

Exemplo 4: Este exemplo descreve as ações de escala programadas que terminam no horário especificado.

```
Get-ASScheduledAction -EndTime "2015-12-30T08:00:00Z"
```

Exemplo 5: Este exemplo descreve as ações de escala programadas para todos os seus grupos do Auto Scaling.

```
Get-ASScheduledAction
```

- Para obter detalhes da API, consulte DescribeScheduledActionsem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DescribeTags`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeTags`.

CLI

AWS CLI

Como descrever todas as tags

Este exemplo descreve todas as tags.

```
aws autoscaling describe-tags
```

Saída:

```
{
    "Tags": [
        {
            "ResourceType": "auto-scaling-group",
            "ResourceId": "my-asg",
            "PropagateAtLaunch": true,
            "Value": "Research",
            "Key": "Dept"
        },
        {
            "ResourceType": "auto-scaling-group",
            "ResourceId": "my-asg",
            "PropagateAtLaunch": true,
            "Value": "WebServer",
            "Key": "Role"
        }
    ]
}
```

Para obter mais informações, consulte Marcar instâncias e grupos de Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: Como descrever tags para um grupo especificado.

Use a opção para descrever as tags de um grupo do Auto Scaling específico use a opção `--filters`.

```
aws autoscaling describe-tags --filters Name=auto-scaling-group,Values=my-asg
```

Para obter mais informações, consulte Marcar instâncias e grupos de Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 3: Para descrever o número específico de tags

Use a opção `--max-items` para retornar um número específico de tags.

```
aws autoscaling describe-tags \
    --max-items 1
```

Se a saída incluir um campo `NextToken`, há mais tags. Para obter os tags adicionais, use o valor desse campo com a opção `--starting-token` em uma chamada subsequente da seguinte maneira.

```
aws autoscaling describe-tags \
    --filters Name=auto-scaling-group,Values=my-asg \
    --starting-token Z3M3LMPEXAMPLE
```

Para obter mais informações, consulte Marcar instâncias e grupos de Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DescribeTagsem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo descreve as tags com um valor-chave de “myTag” ou “myTag2”. Os valores possíveis para o nome do filtro são auto-scaling-group '', 'chave', 'valor' e 'propagate-at-launch'. A sintaxe usada neste exemplo requer a PowerShell versão 3 ou posterior.

```
Get-ASTag -Filter @( @{ Name="key"; Values=@("myTag", "myTag2") } )
```

Saída:

```
Key               : myTag2
PropagateAtLaunch : True
```

```
ResourceId        : my-asg
ResourceType      : auto-scaling-group
Value             : myTagValue2

Key               : myTag
PropagateAtLaunch : True
ResourceId        : my-asg
ResourceType      : auto-scaling-group
Value             : myTagValue
```

Exemplo 2: Com a PowerShell versão 2, você deve usar New-Object para criar o filtro para o parâmetro Filter.

```
$keys = New-Object string[] 2
$keys[0] = "myTag"
$keys[1] = "myTag2"
$filter = New-Object Amazon.AutoScaling.Model.Filter
$filter.Name = "key"
$filter.Values = $keys
Get-ASTag -Filter @( $filter )
```

Exemplo 3: Este exemplo descreve todas as tags para tudos seus grupos do Auto Scaling.

```
Get-ASTag
```

- Para obter detalhes da API, consulte DescribeTagsem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo descreve as tags com um valor-chave de “myTag” ou “myTag2”. Os valores possíveis para o nome do filtro são auto-scaling-group '', 'chave', 'valor' e 'propagate-at-launch'. A sintaxe usada neste exemplo requer a PowerShell versão 3 ou posterior.

```
Get-ASTag -Filter @( @{ Name="key"; Values=@("myTag", "myTag2") } )
```

Saída:

```
Key               : myTag2
PropagateAtLaunch : True
ResourceId        : my-asg
```

```
ResourceType      : auto-scaling-group
Value             : myTagValue2

Key               : myTag
PropagateAtLaunch : True
ResourceId        : my-asg
ResourceType      : auto-scaling-group
Value             : myTagValue
```

Exemplo 2: Com a PowerShell versão 2, você deve usar New-Object para criar o filtro para o parâmetro Filter.

```
$keys = New-Object string[] 2
$keys[0] = "myTag"
$keys[1] = "myTag2"
$filter = New-Object Amazon.AutoScaling.Model.Filter
$filter.Name = "key"
$filter.Values = $keys
Get-ASTag -Filter @( $filter )
```

Exemplo 3: Este exemplo descreve todas as tags para tudos seus grupos do Auto Scaling.

```
Get-ASTag
```

- Para obter detalhes da API, consulte DescribeTagsem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar `DescribeTerminationPolicyTypes` com uma CLI

Os exemplos de código a seguir mostram como usar o `DescribeTerminationPolicyTypes`.

CLI

AWS CLI

Como descrever os tipos de políticas de encerramento disponíveis

Este exemplo descreve os tipos de políticas de encerramento disponíveis.

```
aws autoscaling describe-termination-policy-types
```

Saída:

```
{
    "TerminationPolicyTypes": [
        "AllocationStrategy",
        "ClosestToNextInstanceHour",
        "Default",
        "NewestInstance",
        "OldestInstance",
        "OldestLaunchConfiguration",
        "OldestLaunchTemplate"
    ]
}
```

Para obter mais informações, consulte Controle de quais instâncias do Auto Scaling são terminadas durante a redução da escala horizontalmente no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DescribeTerminationPolicyTypesem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: esse exemplo lista as políticas de terminação que são compatíveis com o Auto Scaling.

```
Get-ASTerminationPolicyType
```

Saída:

```
ClosestToNextInstanceHour
Default
NewestInstance
OldestInstance
```

```
OldestLaunchConfiguration
```

- Para obter detalhes da API, consulte DescribeTerminationPolicyTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: esse exemplo lista as políticas de terminação que são compatíveis com o Auto Scaling.

```
Get-ASTerminationPolicyType
```

Saída:

```
ClosestToNextInstanceHour
Default
NewestInstance
OldestInstance
OldestLaunchConfiguration
```

- Para obter detalhes da API, consulte DescribeTerminationPolicyTypesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DetachInstances`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DetachInstances`.

CLI

AWS CLI

Para desvincular uma instância de um grupo do Auto Scaling existente

Este exemplo desvincula a instância especificada do grupo do Auto Scaling especificado.

```
aws autoscaling detach-instances \
```

```
    --instance-ids i-030017cfa84b20135 \
    --auto-scaling-group-name my-asg \
    --should-decrement-desired-capacity
```

Saída:

```
{
    "Activities": [
        {
            "ActivityId": "5091cb52-547a-47ce-a236-c9ccbc2cb2c9",
            "AutoScalingGroupName": "my-asg",
            "Description": "Detaching EC2 instance: i-030017cfa84b20135",
            "Cause": "At 2020-10-31T17:35:04Z instance i-030017cfa84b20135 was
 detached in response to a user request, shrinking the capacity from 2 to 1.",
            "StartTime": "2020-04-12T15:02:16.179Z",
            "StatusCode": "InProgress",
            "Progress": 50,
            "Details": "{\"Subnet ID\":\"subnet-6194ea3b\",\"Availability Zone\":
\"us-west-2c\"}"
        }
    ]
}
```

- Para obter detalhes da API, consulte DetachInstancesem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo desvincula a instância especificada do grupo do Auto Scaling especificado e diminui a capacidade desejada para que o Auto Scaling não inicie uma instância substituta.

```
Dismount-ASInstance -InstanceId i-93633f9b -AutoScalingGroupName my-asg -
ShouldDecrementDesiredCapacity $true
```

Saída:

```
ActivityId           : 06733445-ce94-4039-be1b-b9f1866e276e
AutoScalingGroupName : my-asg
```

```
Cause                : At 2015-11-20T22:34:59Z instance i-93633f9b was detached
 in response to a user request, shrinking
                       the capacity from 2 to 1.
Description          : Detaching EC2 instance: i-93633f9b
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 50
StartTime            : 11/20/2015 2:34:59 PM
StatusCode           : InProgress
StatusMessage        :
```

Exemplo 2: Este exemplo desvincula a instância especificada do grupo de Auto Scaling especificado sem diminuir a capacidade desejada. O Auto Scaling inicia uma nova instância de substituição.

```
Dismount-ASInstance -InstanceId i-7bf746a2 -AutoScalingGroupName my-asg -
ShouldDecrementDesiredCapacity $false
```

Saída:

```
ActivityId           : f43a3cd4-d38c-4af7-9fe0-d76ec2307b6d
AutoScalingGroupName : my-asg
Cause                : At 2015-11-20T22:34:59Z instance i-7bf746a2 was detached
 in response to a user request.
Description          : Detaching EC2 instance: i-7bf746a2
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 50
StartTime            : 11/20/2015 2:34:59 PM
StatusCode           : InProgress
StatusMessage        :
```

- Para obter detalhes da API, consulte DetachInstancesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo desvincula a instância especificada do grupo do Auto Scaling especificado e diminui a capacidade desejada para que o Auto Scaling não inicie uma instância substituta.

```
Dismount-ASInstance -InstanceId i-93633f9b -AutoScalingGroupName my-asg -
ShouldDecrementDesiredCapacity $true
```

Saída:

```
ActivityId           : 06733445-ce94-4039-be1b-b9f1866e276e
AutoScalingGroupName : my-asg
Cause                : At 2015-11-20T22:34:59Z instance i-93633f9b was detached
 in response to a user request, shrinking
                       the capacity from 2 to 1.
Description          : Detaching EC2 instance: i-93633f9b
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 50
StartTime            : 11/20/2015 2:34:59 PM
StatusCode           : InProgress
StatusMessage        :
```

Exemplo 2: Este exemplo desvincula a instância especificada do grupo de Auto Scaling especificado sem diminuir a capacidade desejada. O Auto Scaling inicia uma nova instância de substituição.

```
Dismount-ASInstance -InstanceId i-7bf746a2 -AutoScalingGroupName my-asg -
ShouldDecrementDesiredCapacity $false
```

Saída:

```
ActivityId           : f43a3cd4-d38c-4af7-9fe0-d76ec2307b6d
AutoScalingGroupName : my-asg
Cause                : At 2015-11-20T22:34:59Z instance i-7bf746a2 was detached
 in response to a user request.
Description          : Detaching EC2 instance: i-7bf746a2
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 50
StartTime            : 11/20/2015 2:34:59 PM
StatusCode           : InProgress
StatusMessage        :
```

- Para obter detalhes da API, consulte DetachInstancesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`DetachLoadBalancers`** com uma CLI

Os exemplos de código a seguir mostram como usar o `DetachLoadBalancers`.

CLI

AWS CLI

Como desvincula um Classic Load Balancer de um grupo do Auto Scaling

Este exemplo desvincula o Classic Load Balancer especificado do grupo do Auto Scaling especificado.

```
aws autoscaling detach-load-balancers \
    --load-balancer-names my-load-balancer \
    --auto-scaling-group-name my-asg
```

Este comando não produz saída.

Para obter mais informações, consulte Anexar um balanceador de carga ao seu grupo Auto Scaling no Guia do Usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DetachLoadBalancersem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo desvincula o balanceador de carga especificado do grupo de Auto Scaling especificado.

```
Dismount-ASLoadBalancer -LoadBalancerName my-lb -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte DetachLoadBalancersem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo desvincula o balanceador de carga especificado do grupo de Auto Scaling especificado.

```
Dismount-ASLoadBalancer -LoadBalancerName my-lb -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte DetachLoadBalancersem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Use **`DisableMetricsCollection`** com um AWS SDK ou CLI

Os exemplos de código a seguir mostram como usar o `DisableMetricsCollection`.

Exemplos de ações são trechos de código de programas maiores e devem ser executados em contexto. É possível ver essa ação em contexto no seguinte exemplo de código:

- Conheça os conceitos básicos

.NET

SDK para .NET (v4)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
/// <summary>
/// Disable the collection of metric data for an Amazon EC2 Auto Scaling
/// group.
/// </summary>
```

```
    /// <param name="groupName">The name of the Auto Scaling group.</param>
    /// <returns>A Boolean value that indicates the success or failure of
    /// the operation.</returns>
    public async Task<bool> DisableMetricsCollectionAsync(string groupName)
    {
        var request = new DisableMetricsCollectionRequest
        {
            AutoScalingGroupName = groupName,
        };

        var response = await
_amazonAutoScaling.DisableMetricsCollectionAsync(request);
        return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
    }
```

- Para obter detalhes da API, consulte DisableMetricsCollectiona Referência AWS SDK para .NET da API.

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

        Aws::AutoScaling::Model::DisableMetricsCollectionRequest request;
        request.SetAutoScalingGroupName(groupName);

        Aws::AutoScaling::Model::DisableMetricsCollectionOutcome outcome =
                autoScalingClient.DisableMetricsCollection(request);
```

```
        if (outcome.IsSuccess()) {
            std::cout << "Metrics collection has been disabled." << std::endl;
        }
        else {
            std::cerr << "Error with AutoScaling::DisableMetricsCollection. "
                      << outcome.GetError().GetMessage()
                      << std::endl;

        }
```

- Para obter detalhes da API, consulte DisableMetricsCollectiona Referência AWS SDK para C++ da API.

CLI

AWS CLI

Como desabilitar a coleta de métricas de um grupo do Auto Scaling

Este exemplo desativa a coleta da métrica `GroupDesiredCapacity` do grupo do Auto Scaling especificado.

```
aws autoscaling disable-metrics-collection \
    --auto-scaling-group-name my-asg \
    --metrics GroupDesiredCapacity
```

Este comando não produz saída.

Para obter mais informações, consulte CloudWatch Métricas de monitoramento para seus grupos e instâncias do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DisableMetricsCollectionem Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
   public static void disableMetricsCollection(AutoScalingClient
autoScalingClient, String groupName) {
       try {
           DisableMetricsCollectionRequest disableMetricsCollectionRequest =
DisableMetricsCollectionRequest.builder()
                   .autoScalingGroupName(groupName)
                   .metrics("GroupMaxSize")
                   .build();


autoScalingClient.disableMetricsCollection(disableMetricsCollectionRequest);
           System.out.println("The disable metrics collection operation was
successful");

       } catch (AutoScalingException e) {
           System.err.println(e.awsErrorDetails().errorMessage());
           System.exit(1);
       }
   }
```

- Para obter detalhes da API, consulte DisableMetricsCollectiona Referência AWS SDK for Java 2.x da API.

Kotlin

SDK para Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun disableMetricsCollection(groupName: String) {
    val disableMetricsCollectionRequest =
        DisableMetricsCollectionRequest {
            autoScalingGroupName = groupName
            metrics = listOf("GroupMaxSize")
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->

 autoScalingClient.disableMetricsCollection(disableMetricsCollectionRequest)
        println("The disable metrics collection operation was successful")
    }
}
```

- Para obter detalhes da API, consulte a DisableMetricsCollectionreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function disableMetricsCollection($autoScalingGroupName)
    {
```

```
        return $this->autoScalingClient->disableMetricsCollection([
            'AutoScalingGroupName' => $autoScalingGroupName,
        ]);
    }
```

- Para obter detalhes da API, consulte DisableMetricsCollectiona Referência AWS SDK para PHP da API.

PowerShell

Ferramentas para PowerShell V4

Example 1: Desativa o monitoramento de métricas especificadas para o grupo de Auto Scaling.

```
Disable-ASMetricsCollection -AutoScalingGroupName my-asg -Metric
 @("GroupMinSize", "GroupMaxSize")
```

Exemplo 2: Esse exemplo desativa o monitoramento de todas as métricas do grupo do Auto Scaling especificado.

```
Disable-ASMetricsCollection -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte DisableMetricsCollectionem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Example 1: Desativa o monitoramento de métricas especificadas para o grupo de Auto Scaling.

```
Disable-ASMetricsCollection -AutoScalingGroupName my-asg -Metric
 @("GroupMinSize", "GroupMaxSize")
```

Exemplo 2: Esse exemplo desativa o monitoramento de todas as métricas do grupo do Auto Scaling especificado.

```
Disable-ASMetricsCollection -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte DisableMetricsCollectionem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def disable_metrics(self, group_name: str) -> Dict[str, Any]:
        """
        Stops CloudWatch metric collection for the Auto Scaling group.

        :param group_name: The name of the group.
        :return: A dictionary with the response from disabling the metrics
 collection.
        :raises ClientError: If there is an error disabling metrics collection.
        """
        try:
            response = self.autoscaling_client.disable_metrics_collection(
                AutoScalingGroupName=group_name
            )
            logger.info(
                f"Successfully disabled metrics collection for group
 '{group_name}'."
            )
            return response
```

```
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            logger.error(
                f"Couldn't disable metrics for group '{group_name}'. Error code:
{error_code}, Message: {err.response['Error']['Message']}"
            )

            if error_code == "ResourceContentionFault":
                logger.error(
                    f"There is a conflict with another operation that is
modifying the Auto Scaling group '{group_name}'. "
                    "Please try again later."
                )
            raise
```

- Para obter detalhes da API, consulte a DisableMetricsCollectionReferência da API AWS SDK for Python (Boto3).

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        // If this fails it's fine, just means there are extra cloudwatch metrics
events for the scale-down.
        let _ = self
            .autoscaling
            .disable_metrics_collection()
            .auto_scaling_group_name(self.auto_scaling_group_name.clone())
            .send()
            .await;
```

- Para obter detalhes da API, consulte a DisableMetricsCollectionreferência da API AWS SDK for Rust.

SAP ABAP

SDK para SAP ABAP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
" Example: iv_group_name = 'my-auto-scaling-group'

TRY.
    ao_asc->disablemetricscollection(
      iv_autoscalinggroupname = iv_group_name ).

    MESSAGE 'Metrics collection disabled successfully' TYPE 'I'.

  CATCH /aws1/cx_ascresrccontionfault INTO DATA(lo_contention).
    RAISE EXCEPTION lo_contention.
  CATCH /aws1/cx_rt_generic INTO DATA(lo_generic_exception).
    RAISE EXCEPTION lo_generic_exception.
ENDTRY.
```

- Para obter detalhes da API, consulte a DisableMetricsCollectionreferência da API AWS SDK for SAP ABAP.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Use **`EnableMetricsCollection`** com um AWS SDK ou CLI

Os exemplos de código a seguir mostram como usar o `EnableMetricsCollection`.

Exemplos de ações são trechos de código de programas maiores e devem ser executados em contexto. É possível ver essa ação em contexto no seguinte exemplo de código:

- Conheça os conceitos básicos

.NET

SDK para .NET (v4)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    /// <summary>
    /// Enable the collection of metric data for an Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Auto Scaling group.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> EnableMetricsCollectionAsync(string groupName)
    {
        var listMetrics = new List<string>
            {
                "GroupMaxSize",
            };

        var collectionRequest = new EnableMetricsCollectionRequest
        {
            AutoScalingGroupName = groupName,
            Metrics = listMetrics,
            Granularity = "1Minute",
        };

        var response = await
_amazonAutoScaling.EnableMetricsCollectionAsync(collectionRequest);
        return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
    }
```

- Para obter detalhes da API, consulte EnableMetricsCollectiona Referência AWS SDK para .NET da API.

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

        Aws::AutoScaling::Model::EnableMetricsCollectionRequest request;
        request.SetAutoScalingGroupName(groupName);

        request.AddMetrics("GroupMinSize");
        request.AddMetrics("GroupMaxSize");
        request.AddMetrics("GroupDesiredCapacity");
        request.AddMetrics("GroupInServiceInstances");
        request.AddMetrics("GroupTotalInstances");
        request.SetGranularity("1Minute");

        Aws::AutoScaling::Model::EnableMetricsCollectionOutcome outcome =
                autoScalingClient.EnableMetricsCollection(request);
        if (outcome.IsSuccess()) {
            std::cout << "Auto Scaling metrics have been enabled."
                      << std::endl;
        }
        else {
            std::cerr << "Error with AutoScaling::EnableMetricsCollection. "
                      << outcome.GetError().GetMessage()
                      << std::endl;

        }
```

- Para obter detalhes da API, consulte EnableMetricsCollectiona Referência AWS SDK para C++ da API.

CLI

AWS CLI

Exemplo 1: como habilitar a coleta de métricas do grupo do Auto Scaling

Este exemplo habilita a coleta de dados do grupo do Auto Scaling especificado.

```
aws autoscaling enable-metrics-collection \
    --auto-scaling-group-name my-asg \
    --granularity "1Minute"
```

Este comando não produz saída.

Para obter mais informações, consulte CloudWatch Métricas de monitoramento para seus grupos e instâncias do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: como coletar dados da métrica especificada de um grupo de Auto Scaling

Use a opção `--metrics` para coletar dados de uma métrica específica.

```
aws autoscaling enable-metrics-collection \
    --auto-scaling-group-name my-asg \
    --metrics GroupDesiredCapacity --granularity "1Minute"
```

Este comando não produz saída.

Para obter mais informações, consulte CloudWatch Métricas de monitoramento para seus grupos e instâncias do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte EnableMetricsCollectionem Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
   public static void enableMetricsCollection(AutoScalingClient
autoScalingClient, String groupName) {
       try {
           EnableMetricsCollectionRequest collectionRequest =
EnableMetricsCollectionRequest.builder()
                   .autoScalingGroupName(groupName)
                   .metrics("GroupMaxSize")
                   .granularity("1Minute")
                   .build();

           autoScalingClient.enableMetricsCollection(collectionRequest);
           System.out.println("The enable metrics collection operation was
successful");

       } catch (AutoScalingException e) {
           System.err.println(e.awsErrorDetails().errorMessage());
           System.exit(1);
       }
   }
```

- Para obter detalhes da API, consulte EnableMetricsCollectiona Referência AWS SDK for Java 2.x da API.

Kotlin

SDK para Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun enableMetricsCollection(groupName: String?) {
    val collectionRequest =
        EnableMetricsCollectionRequest {
            autoScalingGroupName = groupName
            metrics = listOf("GroupMaxSize")
            granularity = "1Minute"
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.enableMetricsCollection(collectionRequest)
        println("The enable metrics collection operation was successful")
    }
}
```

- Para obter detalhes da API, consulte a EnableMetricsCollectionreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function enableMetricsCollection($autoScalingGroupName, $granularity)
    {
```

```
        return $this->autoScalingClient->enableMetricsCollection([
            'AutoScalingGroupName' => $autoScalingGroupName,
            'Granularity' => $granularity,
        ]);
    }
```

- Para obter detalhes da API, consulte EnableMetricsCollectiona Referência AWS SDK para PHP da API.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Ativa o monitoramento de métricas especificadas para o grupo de Auto Scaling.

```
Enable-ASMetricsCollection  -Metric @("GroupMinSize", "GroupMaxSize") -
AutoScalingGroupName my-asg -Granularity 1Minute
```

Exemplo 2: Este exemplo permite o monitoramento de todas as métricas do grupo do Auto Scaling especificado.

```
Enable-ASMetricsCollection -AutoScalingGroupName my-asg -Granularity 1Minute
```

- Para obter detalhes da API, consulte EnableMetricsCollectionem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Ativa o monitoramento de métricas especificadas para o grupo de Auto Scaling.

```
Enable-ASMetricsCollection  -Metric @("GroupMinSize", "GroupMaxSize") -
AutoScalingGroupName my-asg -Granularity 1Minute
```

Exemplo 2: Este exemplo permite o monitoramento de todas as métricas do grupo do Auto Scaling especificado.

```
Enable-ASMetricsCollection -AutoScalingGroupName my-asg -Granularity 1Minute
```

- Para obter detalhes da API, consulte EnableMetricsCollectionem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def enable_metrics(self, group_name: str, metrics: List[str]) -> Dict[str,
 Any]:
        """
        Enables CloudWatch metric collection for Amazon EC2 Auto Scaling
 activities.

        :param group_name: The name of the group to enable.
        :param metrics: A list of metrics to collect.
        :return: A dictionary with the response from enabling the metrics
 collection.
        :raises ClientError: If there is an error enabling metrics collection.
        """
        try:
            response = self.autoscaling_client.enable_metrics_collection(
                AutoScalingGroupName=group_name, Metrics=metrics,
 Granularity="1Minute"
            )
            logger.info(
                f"Successfully enabled metrics for Auto Scaling group
 '{group_name}'."
            )
```

```
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            logger.error(
                f"Couldn't enable metrics on '{group_name}'. Error code:
{error_code}, Message: {err.response['Error']['Message']}"
            )

            if error_code == "ResourceContentionFault":
                logger.error(
                    f"There is a conflict with another operation that is
modifying the Auto Scaling group '{group_name}'. "
                    "Please try again later."
                )
            elif error_code == "InvalidParameterCombination":
                logger.error(
                    f"The combination of parameters provided for enabling metrics
on '{group_name}' is not valid. "
                    "Please check the parameters and try again."
                )
            raise
        else:
            return response
```

- Para obter detalhes da API, consulte a EnableMetricsCollectionReferência da API AWS SDK for Python (Boto3).

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        let enable_metrics_collection = autoscaling
            .enable_metrics_collection()
            .auto_scaling_group_name(auto_scaling_group_name.as_str())
```

```
            .granularity("1Minute")
            .set_metrics(Some(vec![
                String::from("GroupMinSize"),
                String::from("GroupMaxSize"),
                String::from("GroupDesiredCapacity"),
                String::from("GroupInServiceInstances"),
                String::from("GroupTotalInstances"),
            ]))
            .send()
            .await;
```

- Para obter detalhes da API, consulte a EnableMetricsCollectionreferência da API AWS SDK for Rust.

SAP ABAP

SDK para SAP ABAP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
   " Example: iv_group_name = 'my-auto-scaling-group'
   " Example: it_metrics contains list of metrics like 'GroupMinSize',
'GroupMaxSize', etc.

   TRY.
       ao_asc->enablemetricscollection(
         iv_autoscalinggroupname = iv_group_name
         it_metrics = it_metrics
         iv_granularity = '1Minute' ).

       MESSAGE 'Metrics collection enabled successfully' TYPE 'I'.

     CATCH /aws1/cx_ascresrccontionfault INTO DATA(lo_contention).
       RAISE EXCEPTION lo_contention.
     CATCH /aws1/cx_rt_generic INTO DATA(lo_generic_exception).
       RAISE EXCEPTION lo_generic_exception.
```

```
    ENDTRY.
```

- Para obter detalhes da API, consulte a EnableMetricsCollectionreferência da API AWS SDK for SAP ABAP.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`EnterStandby`** com uma CLI

Os exemplos de código a seguir mostram como usar o `EnterStandby`.

CLI

AWS CLI

Como colocar instâncias no modo de espera

Este exemplo coloca a instância especificada no modo de espera. Isso é útil para atualizar ou solucionar problemas de uma instância que está atualmente em serviço.

```
aws autoscaling enter-standby \
    --instance-ids i-061c63c5eb45f0416 \
    --auto-scaling-group-name my-asg \
    --should-decrement-desired-capacity
```

Saída:

```
{
    "Activities": [
        {
            "ActivityId": "ffa056b4-6ed3-41ba-ae7c-249dfae6eba1",
            "AutoScalingGroupName": "my-asg",
            "Description": "Moving EC2 instance to Standby: i-061c63c5eb45f0416",
            "Cause": "At 2020-10-31T20:31:00Z instance i-061c63c5eb45f0416 was
 moved to standby in response to a user request, shrinking the capacity from 1 to
 0.",
            "StartTime": "2020-10-31T20:31:00.949Z",
```

```
            "StatusCode": "InProgress",
            "Progress": 50,
            "Details": "{\"Subnet ID\":\"subnet-6194ea3b\",\"Availability Zone\":
\"us-west-2c\"}"
        }
    ]
}
```

Para obter mais informações, consulte Amazon EC2 Auto Scaling instância de ciclo de vida no Manual do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte EnterStandbyem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo coloca a instância especificada em modo de espera e diminui a capacidade desejada para que o Auto Scaling não inicie uma instância substituta.

```
Enter-ASStandby -InstanceId i-93633f9b -AutoScalingGroupName my-asg -
ShouldDecrementDesiredCapacity $true
```

Saída:

```
ActivityId           : e36a5a54-ced6-4df8-bd19-708e2a59a649
AutoScalingGroupName : my-asg
Cause                : At 2015-11-22T15:48:06Z instance i-95b8484f was moved to
 standby in response to a user request,
                       shrinking the capacity from 2 to 1.
Description          : Moving EC2 instance to Standby: i-95b8484f
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 50
StartTime            : 11/22/2015 7:48:06 AM
StatusCode           : InProgress
StatusMessage        :
```

Exemplo 2: Este exemplo coloca a instância especificada em modo de espera sem diminuir a capacidade desejada. O Auto Scaling inicia uma nova instância de substituição.

```
Enter-ASStandby -InstanceId i-93633f9b -AutoScalingGroupName my-asg -
ShouldDecrementDesiredCapacity $false
```

Saída:

```
ActivityId           : e36a5a54-ced6-4df8-bd19-708e2a59a649
AutoScalingGroupName : my-asg
Cause                : At 2015-11-22T15:48:06Z instance i-95b8484f was moved to
 standby in response to a user request.
Description          : Moving EC2 instance to Standby: i-95b8484f
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 50
StartTime            : 11/22/2015 7:48:06 AM
StatusCode           : InProgress
StatusMessage        :
```

- Para obter detalhes da API, consulte EnterStandbyem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo coloca a instância especificada em modo de espera e diminui a capacidade desejada para que o Auto Scaling não inicie uma instância substituta.

```
Enter-ASStandby -InstanceId i-93633f9b -AutoScalingGroupName my-asg -
ShouldDecrementDesiredCapacity $true
```

Saída:

```
ActivityId           : e36a5a54-ced6-4df8-bd19-708e2a59a649
AutoScalingGroupName : my-asg
Cause                : At 2015-11-22T15:48:06Z instance i-95b8484f was moved to
 standby in response to a user request,
                       shrinking the capacity from 2 to 1.
Description          : Moving EC2 instance to Standby: i-95b8484f
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 50
StartTime            : 11/22/2015 7:48:06 AM
```

```
StatusCode           : InProgress
StatusMessage        :
```

Exemplo 2: Este exemplo coloca a instância especificada em modo de espera sem diminuir a capacidade desejada. O Auto Scaling inicia uma nova instância de substituição.

```
Enter-ASStandby -InstanceId i-93633f9b -AutoScalingGroupName my-asg -
ShouldDecrementDesiredCapacity $false
```

Saída:

```
ActivityId           : e36a5a54-ced6-4df8-bd19-708e2a59a649
AutoScalingGroupName : my-asg
Cause                : At 2015-11-22T15:48:06Z instance i-95b8484f was moved to
 standby in response to a user request.
Description          : Moving EC2 instance to Standby: i-95b8484f
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 50
StartTime            : 11/22/2015 7:48:06 AM
StatusCode           : InProgress
StatusMessage        :
```

- Para obter detalhes da API, consulte EnterStandbyem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`ExecutePolicy`** com uma CLI

Os exemplos de código a seguir mostram como usar o `ExecutePolicy`.

CLI

AWS CLI

Como executar uma política de escalabilidade

Este exemplo executa a política de escalabilidade chamada `my-step-scale-out-policy` para o grupo do Auto Scaling especificado.

```
aws autoscaling execute-policy \
    --auto-scaling-group-name my-asg \
    --policy-name  my-step-scale-out-policy \
    --metric-value 95 \
    --breach-threshold 80
```

Este comando não produz saída.

Para mais informações, consulte as políticas de escalonamento por passos e escalonamento simples no Guia do Usuário de Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte ExecutePolicyem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo executa a política especificada para o grupo do Auto Scaling especificado.

```
Start-ASPolicy -AutoScalingGroupName my-asg -PolicyName "myScaleInPolicy"
```

Exemplo 2: Este exemplo executa a política especificada para o grupo do Auto Scaling especificado, depois de aguardar a conclusão do período de espera.

```
Start-ASPolicy -AutoScalingGroupName my-asg -PolicyName "myScaleInPolicy" -
HonorCooldown $true
```

- Para obter detalhes da API, consulte ExecutePolicyem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo executa a política especificada para o grupo do Auto Scaling especificado.

```
Start-ASPolicy -AutoScalingGroupName my-asg -PolicyName "myScaleInPolicy"
```

Exemplo 2: Este exemplo executa a política especificada para o grupo do Auto Scaling especificado, depois de aguardar a conclusão do período de espera.

```
Start-ASPolicy -AutoScalingGroupName my-asg -PolicyName "myScaleInPolicy" -
HonorCooldown $true
```

- Para obter detalhes da API, consulte ExecutePolicyem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`ExitStandby`** com uma CLI

Os exemplos de código a seguir mostram como usar o `ExitStandby`.

CLI

AWS CLI

Como tirar instâncias do modo de espera

Esse exemplo move a instância especificada para fora do modo em espera.

```
aws autoscaling exit-standby \
    --instance-ids i-061c63c5eb45f0416 \
    --auto-scaling-group-name my-asg
```

Saída:

```
{
    "Activities": [
        {
            "ActivityId": "142928e1-a2dc-453a-9b24-b85ad6735928",
            "AutoScalingGroupName": "my-asg",
            "Description": "Moving EC2 instance out of Standby:
 i-061c63c5eb45f0416",
            "Cause": "At 2020-10-31T20:32:50Z instance i-061c63c5eb45f0416 was
 moved out of standby in response to a user request, increasing the capacity from
 0 to 1.",
            "StartTime": "2020-10-31T20:32:50.222Z",
```

```
            "StatusCode": "PreInService",
            "Progress": 30,
            "Details": "{\"Subnet ID\":\"subnet-6194ea3b\",\"Availability Zone\":
\"us-west-2c\"}"
        }
    ]
}
```

Para obter mais informações, consulte Desvincular temporariamente instâncias do grupo do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte ExitStandbyem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo tira as instâncias especificadas do modo de espera.

```
Exit-ASStandby -InstanceId i-93633f9b -AutoScalingGroupName my-asg
```

Saída:

```
ActivityId           : 1833d3e8-e32f-454e-b731-0670ad4c6934
AutoScalingGroupName : my-asg
Cause                : At 2015-11-22T15:51:21Z instance i-95b8484f was moved out
 of standby in response to a user
                       request, increasing the capacity from 1 to 2.
Description          : Moving EC2 instance out of Standby: i-95b8484f
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 30
StartTime            : 11/22/2015 7:51:21 AM
StatusCode           : PreInService
StatusMessage        :
```

- Para obter detalhes da API, consulte ExitStandbyem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo tira as instâncias especificadas do modo de espera.

```
Exit-ASStandby -InstanceId i-93633f9b -AutoScalingGroupName my-asg
```

Saída:

```
ActivityId           : 1833d3e8-e32f-454e-b731-0670ad4c6934
AutoScalingGroupName : my-asg
Cause                : At 2015-11-22T15:51:21Z instance i-95b8484f was moved out
 of standby in response to a user
                       request, increasing the capacity from 1 to 2.
Description          : Moving EC2 instance out of Standby: i-95b8484f
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 30
StartTime            : 11/22/2015 7:51:21 AM
StatusCode           : PreInService
StatusMessage        :
```

- Para obter detalhes da API, consulte ExitStandbyem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`PutLifecycleHook`** com uma CLI

Os exemplos de código a seguir mostram como usar o `PutLifecycleHook`.

CLI

AWS CLI

Exemplo 1: Como criar um gancho do ciclo de vida

Este exemplo cria um gancho do ciclo de vida que será invocado em qualquer instância recém-iniciada, com um tempo limite de 4800 segundos. Isso é útil para manter as instâncias em um estado de espera até que os scripts de dados do usuário terminem ou para invocar uma função AWS Lambda usando. EventBridge

```
aws autoscaling put-lifecycle-hook \
```

```
    --auto-scaling-group-name my-asg \
    --lifecycle-hook-name my-launch-hook \
    --lifecycle-transition autoscaling:EC2_INSTANCE_LAUNCHING \
    --heartbeat-timeout 4800
```

Este comando não produz saída. Se já existir um gancho do ciclo de vida com o mesmo nome, ele será substituído pelo novo gancho do ciclo de vida.

Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: Como enviar uma mensagem de e-mail do Amazon SNS para notificar sobre as transições de estado da instância.

Este exemplo cria um gancho do ciclo de vida com o tópico do Amazon SNS e o perfil do IAM a serem usados para receber notificações na execução da instância.

```
aws autoscaling put-lifecycle-hook \
    --auto-scaling-group-name my-asg \
    --lifecycle-hook-name my-launch-hook \
    --lifecycle-transition autoscaling:EC2_INSTANCE_LAUNCHING \
    --notification-target-arn arn:aws:sns:us-west-2:123456789012:my-sns-topic \
    --role-arn arn:aws:iam::123456789012:role/my-auto-scaling-role
```

Este comando não produz saída.

Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 3: Como publicar uma mensagem em uma fila do Amazon SQS.

Este exemplo cria um gancho do ciclo de vida que publica uma mensagem com metadados na fila especificada do Amazon SQS.

```
aws autoscaling put-lifecycle-hook \
    --auto-scaling-group-name my-asg \
    --lifecycle-hook-name my-launch-hook \
    --lifecycle-transition autoscaling:EC2_INSTANCE_LAUNCHING \
    --notification-target-arn arn:aws:sqs:us-west-2:123456789012:my-sqs-queue \
    --role-arn arn:aws:iam::123456789012:role/my-notification-role \
    --notification-metadata "SQS message metadata"
```

Este comando não produz saída.

Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte PutLifecycleHookem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo adiciona o gancho do ciclo de vida especificado ao grupo do Auto Scaling especificado.

```
Write-ASLifecycleHook -AutoScalingGroupName my-asg -LifecycleHookName
 "myLifecycleHook" -LifecycleTransition "autoscaling:EC2_INSTANCE_LAUNCHING" -
NotificationTargetARN "arn:aws:sns:us-west-2:123456789012:my-sns-topic" -RoleARN
 "arn:aws:iam::123456789012:role/my-iam-role"
```

- Para obter detalhes da API, consulte PutLifecycleHookem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo adiciona o gancho do ciclo de vida especificado ao grupo do Auto Scaling especificado.

```
Write-ASLifecycleHook -AutoScalingGroupName my-asg -LifecycleHookName
 "myLifecycleHook" -LifecycleTransition "autoscaling:EC2_INSTANCE_LAUNCHING" -
NotificationTargetARN "arn:aws:sns:us-west-2:123456789012:my-sns-topic" -RoleARN
 "arn:aws:iam::123456789012:role/my-iam-role"
```

- Para obter detalhes da API, consulte PutLifecycleHookem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **PutNotificationConfiguration** com uma CLI

Os exemplos de código a seguir mostram como usar o `PutNotificationConfiguration`.

CLI

AWS CLI

Para adicionar uma nova notificação

Este exemplo adiciona o grupo de destino especificado ao grupo do Auto Scaling especificado.

```
aws autoscaling put-notification-configuration \
    --auto-scaling-group-name my-asg \
    --topic-arn arn:aws:sns:us-west-2:123456789012:my-sns-topic \
    --notification-type autoscaling:TEST_NOTIFICATION
```

Este comando não produz saída.

Para obter mais informações, consulte Obter notificações do Amazon SNS quando o grupo de Auto Scaling é dimensionado no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte PutNotificationConfigurationem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo configura o grupo do Auto Scaling especificado para enviar uma notificação ao tópico SNS especificado quando ele iniciar instâncias do EC2.

```
Write-ASNotificationConfiguration -AutoScalingGroupName my-asg -
NotificationType "autoscaling:EC2_INSTANCE_LAUNCH" -TopicARN "arn:aws:sns:us-
west-2:123456789012:my-topic"
```

Exemplo 2: Este exemplo configura o grupo do Auto Scaling especificado para enviar uma notificação ao tópico SNS especificado quando ele iniciar ou encerrar instâncias do EC2.

```
Write-ASNotificationConfiguration -AutoScalingGroupName my-asg -NotificationType
 @("autoscaling:EC2_INSTANCE_LAUNCH", "autoscaling:EC2_INSTANCE_TERMINATE") -
TopicARN "arn:aws:sns:us-west-2:123456789012:my-topic"
```

- Para obter detalhes da API, consulte PutNotificationConfigurationem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo configura o grupo do Auto Scaling especificado para enviar uma notificação ao tópico SNS especificado quando ele iniciar instâncias do EC2.

```
Write-ASNotificationConfiguration -AutoScalingGroupName my-asg -
NotificationType "autoscaling:EC2_INSTANCE_LAUNCH" -TopicARN "arn:aws:sns:us-
west-2:123456789012:my-topic"
```

Exemplo 2: Este exemplo configura o grupo do Auto Scaling especificado para enviar uma notificação ao tópico SNS especificado quando ele iniciar ou encerrar instâncias do EC2.

```
Write-ASNotificationConfiguration -AutoScalingGroupName my-asg -NotificationType
 @("autoscaling:EC2_INSTANCE_LAUNCH", "autoscaling:EC2_INSTANCE_TERMINATE") -
TopicARN "arn:aws:sns:us-west-2:123456789012:my-topic"
```

- Para obter detalhes da API, consulte PutNotificationConfigurationem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`PutScalingPolicy`** com uma CLI

Os exemplos de código a seguir mostram como usar o `PutScalingPolicy`.

CLI

AWS CLI

Para adicionar uma política de escalonamento de acompanhamento de alvo a um grupo de Auto Scaling

O exemplo `put-scaling-policy` a seguir aplica uma política de escalabilidade de monitoramento de destino ao grupo do Auto Scaling especificado. A saída contém os nomes ARNs e dos dois CloudWatch alarmes criados em seu nome. Se já existir uma política de escalonamento com o mesmo nome, ela será substituída pela nova política de escalonamento.

```
aws autoscaling put-scaling-policy --auto-scaling-group-name my-asg \
  --policy-name alb1000-target-tracking-scaling-policy \
  --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
```

Conteúdo de `config.json`:

```
{
     "TargetValue": 1000.0,
     "PredefinedMetricSpecification": {
          "PredefinedMetricType": "ALBRequestCountPerTarget",
          "ResourceLabel": "app/my-alb/778d41231b141a0f/targetgroup/my-alb-
target-group/943f017f100becff"
     }
}
```

Saída:

```
{
     "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:228f02c2-
c665-4bfd-aaac-8b04080bea3c:autoScalingGroupName/my-asg:policyName/alb1000-
target-tracking-scaling-policy",
     "Alarms": [
         {
             "AlarmARN": "arn:aws:cloudwatch:region:account-
id:alarm:TargetTracking-my-asg-AlarmHigh-fc0e4183-23ac-497e-9992-691c9980c38e",
             "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
         },
         {
             "AlarmARN": "arn:aws:cloudwatch:region:account-
id:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-bd9e-471a352ee1a2",
             "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
         }
     ]
```

```
}
```

Para obter mais exemplos, consulte Exemplos de políticas de escalabilidade para a interface de linha de AWS comando (AWS CLI) no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte PutScalingPolicyem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo adiciona a política especificada ao grupo do Auto Scaling especificado. O tipo de ajuste especificado determina como interpretar o ScalingAdjustment parâmetro. Com 'ChangeInCapacity', um valor positivo aumenta a capacidade pelo número especificado de instâncias e um valor negativo diminui a capacidade pelo número especificado de instâncias.

```
Write-ASScalingPolicy -AutoScalingGroupName my-asg -AdjustmentType
 "ChangeInCapacity" -PolicyName "myScaleInPolicy" -ScalingAdjustment -1
```

Saída:

```
arn:aws:autoscaling:us-west-2:123456789012:scalingPolicy:aa3836ab-5462-42c7-adab-
e1d769fc24ef:autoScalingGroupName/my-asg
:policyName/myScaleInPolicy
```

- Para obter detalhes da API, consulte PutScalingPolicyem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo adiciona a política especificada ao grupo do Auto Scaling especificado. O tipo de ajuste especificado determina como interpretar o ScalingAdjustment parâmetro. Com 'ChangeInCapacity', um valor positivo aumenta a capacidade pelo número especificado de instâncias e um valor negativo diminui a capacidade pelo número especificado de instâncias.

```
Write-ASScalingPolicy -AutoScalingGroupName my-asg -AdjustmentType
 "ChangeInCapacity" -PolicyName "myScaleInPolicy" -ScalingAdjustment -1
```

Saída:

```
arn:aws:autoscaling:us-west-2:123456789012:scalingPolicy:aa3836ab-5462-42c7-adab-
e1d769fc24ef:autoScalingGroupName/my-asg
:policyName/myScaleInPolicy
```

- Para obter detalhes da API, consulte PutScalingPolicyem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **PutScheduledUpdateGroupAction** com uma CLI

Os exemplos de código a seguir mostram como usar o `PutScheduledUpdateGroupAction`.

CLI

AWS CLI

Exemplo 1: Para adicionar uma ação programada a um grupo do Auto Scaling

Este exemplo adiciona a ação agendada especificada ao grupo do Auto Scaling especificado.

```
aws autoscaling put-scheduled-update-group-action \
    --auto-scaling-group-name my-asg \
    --scheduled-action-name my-scheduled-action \
    --start-time "2023-05-12T08:00:00Z" \
    --min-size 2 \
    --max-size 6 \
    --desired-capacity 4
```

Este comando não produz saída. Se uma ação programada com o mesmo nome já existir, ela será substituída pela nova ação programada.

Para obter mais exemplos, consulte Escalabilidade programada no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: especificar uma programação recorrente

Este exemplo cria uma ação programada para escalar em uma programação recorrente que está programada para ser executada às 00h30 horas nos primeiros dias de janeiro, junho e dezembro de cada ano.

```
aws autoscaling put-scheduled-update-group-action \
    --auto-scaling-group-name my-asg \
    --scheduled-action-name my-recurring-action \
    --recurrence "30 0 1 1,6,12 *" \
    --min-size 2 \
    --max-size 6 \
    --desired-capacity 4
```

Este comando não produz saída. Se uma ação programada com o mesmo nome já existir, ela será substituída pela nova ação programada.

Para obter mais exemplos, consulte Escalabilidade programada no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte PutScheduledUpdateGroupActionem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo cria ou atualiza uma ação programada única para alterar a capacidade desejada na hora de início especificada.

```
Write-ASScheduledUpdateGroupAction -AutoScalingGroupName my-asg -
ScheduledActionName "myScheduledAction" -StartTime "2015-12-01T00:00:00Z" -
DesiredCapacity 10
```

- Para obter detalhes da API, consulte PutScheduledUpdateGroupActionem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo cria ou atualiza uma ação programada única para alterar a capacidade desejada na hora de início especificada.

```
Write-ASScheduledUpdateGroupAction -AutoScalingGroupName my-asg -
ScheduledActionName "myScheduledAction" -StartTime "2015-12-01T00:00:00Z" -
DesiredCapacity 10
```

- Para obter detalhes da API, consulte PutScheduledUpdateGroupActionem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`RecordLifecycleActionHeartbeat`** com uma CLI

Os exemplos de código a seguir mostram como usar o `RecordLifecycleActionHeartbeat`.

CLI

AWS CLI

Como registrar a pulsação de uma ação do ciclo de vida

Este exemplo registra uma pulsação da ação do ciclo de vida para manter a instância em um estado pendente.

```
aws autoscaling record-lifecycle-action-heartbeat \
    --lifecycle-hook-name my-launch-hook \
    --auto-scaling-group-name my-asg \
    --lifecycle-action-token bcd2f1b8-9a78-44d3-8a7a-4dd07d7cf635
```

Este comando não produz saída.

Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte RecordLifecycleActionHeartbeatem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo registra uma pulsação para a ação de ciclo de vida especificada. Isso mantém a instância em um estado pendente até que você conclua a ação personalizada.

```
Write-ASLifecycleActionHeartbeat -AutoScalingGroupName my-asg -LifecycleHookName
 myLifecycleHook -LifecycleActionToken bcd2f1b8-9a78-44d3-8a7a-4dd07d7cf635
```

- Para obter detalhes da API, consulte RecordLifecycleActionHeartbeatem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo registra uma pulsação para a ação de ciclo de vida especificada. Isso mantém a instância em um estado pendente até que você conclua a ação personalizada.

```
Write-ASLifecycleActionHeartbeat -AutoScalingGroupName my-asg -LifecycleHookName
 myLifecycleHook -LifecycleActionToken bcd2f1b8-9a78-44d3-8a7a-4dd07d7cf635
```

- Para obter detalhes da API, consulte RecordLifecycleActionHeartbeatem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`ResumeProcesses`** com uma CLI

Os exemplos de código a seguir mostram como usar o `ResumeProcesses`.

CLI

AWS CLI

Para retomar um processo suspenso

Este exemplo retoma o processo de escalabilidade suspenso especificado para o grupo do Auto Scaling especificado.

```
aws autoscaling resume-processes \
    --auto-scaling-group-name my-asg \
    --scaling-processes AlarmNotification
```

Este comando não produz saída.

Para obter mais informações, consulte Suspensão e retomada dos processos de escalabilidade no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte ResumeProcessesem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Esse exemplo reinicia o processo do Auto Scaling especificado para o grupo de Auto Scaling especificado.

```
Resume-ASProcess -AutoScalingGroupName my-asg -ScalingProcess "AlarmNotification"
```

Exemplo 2: Este exemplo retoma todos os processos suspensos do Auto Scaling para o grupo do Auto Scaling especificado.

```
Resume-ASProcess -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte ResumeProcessesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Esse exemplo reinicia o processo do Auto Scaling especificado para o grupo de Auto Scaling especificado.

```
Resume-ASProcess -AutoScalingGroupName my-asg -ScalingProcess "AlarmNotification"
```

Exemplo 2: Este exemplo retoma todos os processos suspensos do Auto Scaling para o grupo do Auto Scaling especificado.

```
Resume-ASProcess -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte ResumeProcessesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Use **`SetDesiredCapacity`** com um AWS SDK ou CLI

Os exemplos de código a seguir mostram como usar o `SetDesiredCapacity`.

Exemplos de ações são trechos de código de programas maiores e devem ser executados em contexto. É possível ver essa ação em contexto no seguinte exemplo de código:

- Conheça os conceitos básicos

.NET

SDK para .NET (v4)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
/// <summary>
/// Set the desired capacity of an Auto Scaling group.
/// </summary>
/// <param name="groupName">The name of the Auto Scaling group.</param>
/// <param name="desiredCapacity">The desired capacity for the Auto
/// Scaling group.</param>
/// <returns>A Boolean value indicating the success of the action.</returns>
public async Task<bool> SetDesiredCapacityAsync(
    string groupName,
    int desiredCapacity)
{
    var capacityRequest = new SetDesiredCapacityRequest
    {
        AutoScalingGroupName = groupName,
```

```
            DesiredCapacity = desiredCapacity,
        };

        var response = await
_amazonAutoScaling.SetDesiredCapacityAsync(capacityRequest);
        Console.WriteLine($"You have set the DesiredCapacity to
{desiredCapacity}.");

        return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
    }
```

- Para obter detalhes da API, consulte SetDesiredCapacitya Referência AWS SDK para .NET da API.

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

        Aws::AutoScaling::Model::SetDesiredCapacityRequest request;
        request.SetAutoScalingGroupName(groupName);
        request.SetDesiredCapacity(2);

        Aws::AutoScaling::Model::SetDesiredCapacityOutcome outcome =
                autoScalingClient.SetDesiredCapacity(request);

        if (!outcome.IsSuccess()) {
            std::cerr << "Error with AutoScaling::SetDesiredCapacityRequest. "
```

```
                    << outcome.GetError().GetMessage()
                    << std::endl;

        }
```

- Para obter detalhes da API, consulte SetDesiredCapacitya Referência AWS SDK para C++ da API.

CLI

AWS CLI

Como definir a capacidade desejada para um grupo do Auto Scaling

Este exemplo define a capacidade desejada para o grupo do Auto Scaling especificado.

```
aws autoscaling set-desired-capacity \
    --auto-scaling-group-name my-asg \
    --desired-capacity 2 \
    --honor-cooldown
```

Esse comando retornará ao prompt, se houver êxito.

- Para obter detalhes da API, consulte SetDesiredCapacityem Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
   public static void setDesiredCapacity(AutoScalingClient autoScalingClient,
String groupName) {
```

```
        try {
            SetDesiredCapacityRequest capacityRequest =
 SetDesiredCapacityRequest.builder()
                    .autoScalingGroupName(groupName)
                    .desiredCapacity(2)
                    .build();

            autoScalingClient.setDesiredCapacity(capacityRequest);
            System.out.println("You have set the DesiredCapacity to 2");

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
```

- Para obter detalhes da API, consulte SetDesiredCapacitya Referência AWS SDK for Java 2.x da API.

Kotlin

SDK para Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun setDesiredCapacity(groupName: String) {
    val capacityRequest =
        SetDesiredCapacityRequest {
            autoScalingGroupName = groupName
            desiredCapacity = 2
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.setDesiredCapacity(capacityRequest)
        println("You set the DesiredCapacity to 2")
    }
```

```
}
```

- Para obter detalhes da API, consulte a SetDesiredCapacityreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function setDesiredCapacity($autoScalingGroupName, $desiredCapacity)
    {
        return $this->autoScalingClient->setDesiredCapacity([
            'AutoScalingGroupName' => $autoScalingGroupName,
            'DesiredCapacity' => $desiredCapacity,
        ]);
    }
```

- Para obter detalhes da API, consulte SetDesiredCapacitya Referência AWS SDK para PHP da API.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Define o tamanho do grupo de Auto Scaling especificado.

```
Set-ASDesiredCapacity -AutoScalingGroupName my-asg -DesiredCapacity 2
```

Exemplo 2: Este exemplo define o tamanho do grupo do Auto Scaling especificado e aguarda a conclusão do período de espera antes de fazer o escalonamento para o novo tamanho.

```
Set-ASDesiredCapacity -AutoScalingGroupName my-asg -DesiredCapacity 2 -
HonorCooldown $true
```

- Para obter detalhes da API, consulte SetDesiredCapacityem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Define o tamanho do grupo de Auto Scaling especificado.

```
Set-ASDesiredCapacity -AutoScalingGroupName my-asg -DesiredCapacity 2
```

Exemplo 2: Este exemplo define o tamanho do grupo do Auto Scaling especificado e aguarda a conclusão do período de espera antes de fazer o escalonamento para o novo tamanho.

```
Set-ASDesiredCapacity -AutoScalingGroupName my-asg -DesiredCapacity 2 -
HonorCooldown $true
```

- Para obter detalhes da API, consulte SetDesiredCapacityem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client
```

```
    def set_desired_capacity(self, group_name: str, capacity: int) -> None:
        """
        Sets the desired capacity of the group. Amazon EC2 Auto Scaling tries to
keep the
        number of running instances equal to the desired capacity.

        :param group_name: The name of the group to update.
        :param capacity: The desired number of running instances.
        :return: None
        :raises ClientError: If there is an error setting the desired capacity.
        """
        try:
            self.autoscaling_client.set_desired_capacity(
                AutoScalingGroupName=group_name,
                DesiredCapacity=capacity,
                HonorCooldown=False,
            )
            logger.info(
                f"Successfully set desired capacity of {capacity} for Auto
Scaling group '{group_name}'."
            )

        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            logger.error(
                f"Failed to set desired capacity for Auto Scaling group
'{group_name}'."
            )
            if error_code == "ScalingActivityInProgress":
                logger.error(
                    f"A scaling activity is currently in progress for the Auto
Scaling group '{group_name}'. "
                    "Please wait for the activity to complete before attempting
to set the desired capacity."
                )
            logger.error(f"Full error:\n\t{err}")
            raise
```

- Para obter detalhes da API, consulte a SetDesiredCapacityReferência da API AWS SDK for Python (Boto3).

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
   pub async fn scale_desired_capacity(&self, capacity: i32) -> Result<(),
ScenarioError> {
       // 7. SetDesiredCapacity: set desired capacity to 2.
       //   Wait for a second instance to launch.
       let update_group = self
           .autoscaling
           .set_desired_capacity()
           .auto_scaling_group_name(self.auto_scaling_group_name.clone())
           .desired_capacity(capacity)
           .send()
           .await;
       if let Err(err) = update_group {
           return Err(ScenarioError::new(
               format!("Failed to update group to desired capacity
({capacity})").as_str(),
               &err,
           ));
       }
       Ok(())
   }
```

- Para obter detalhes da API, consulte a SetDesiredCapacityreferência da API AWS SDK for Rust.

SAP ABAP

SDK para SAP ABAP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
" Example: iv_group_name = 'my-auto-scaling-group'
" Example: iv_capacity = 2

TRY.
    ao_asc->setdesiredcapacity(
      iv_autoscalinggroupname = iv_group_name
      iv_desiredcapacity = iv_capacity
      iv_honorcooldown = abap_false ).

    MESSAGE 'Desired capacity set successfully' TYPE 'I'.

  CATCH /aws1/cx_ascscaactivityinprg00 INTO DATA(lo_activity_in_progress).
    RAISE EXCEPTION lo_activity_in_progress.
  CATCH /aws1/cx_rt_generic INTO DATA(lo_generic_exception).
    RAISE EXCEPTION lo_generic_exception.
ENDTRY.
```

- Para obter detalhes da API, consulte a SetDesiredCapacityreferência da API AWS SDK for SAP ABAP.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`SetInstanceHealth`** com uma CLI

Os exemplos de código a seguir mostram como usar o `SetInstanceHealth`.

CLI

AWS CLI

Para verificar o status de integridade de uma instância

Define o status da integridade da instância especificada para `Unhealthy`.

```
aws autoscaling set-instance-health \
    --instance-id i-061c63c5eb45f0416 \
    --health-status Unhealthy
```

Este comando não produz saída.

- Para obter detalhes da API, consulte SetInstanceHealthem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo define o status da instância especificada como “Não saudável”, tirando-a de serviço. O Auto Scaling termina e substitui a instância.

```
Set-ASInstanceHealth -HealthStatus Unhealthy -InstanceId i-93633f9b
```

Exemplo 2: Este exemplo define o status da instância especificada como “Saudável”, mantendo-a em serviço. Definir um período de carência da verificação de integridade para um grupo do Auto Scaling

```
Set-ASInstanceHealth -HealthStatus Healthy -InstanceId i-93633f9b -
ShouldRespectGracePeriod $false
```

- Para obter detalhes da API, consulte SetInstanceHealthem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo define o status da instância especificada como “Não saudável”, tirando-a de serviço. O Auto Scaling termina e substitui a instância.

```
Set-ASInstanceHealth -HealthStatus Unhealthy -InstanceId i-93633f9b
```

Exemplo 2: Este exemplo define o status da instância especificada como “Saudável”, mantendo-a em serviço. Definir um período de carência da verificação de integridade para um grupo do Auto Scaling

```
Set-ASInstanceHealth -HealthStatus Healthy -InstanceId i-93633f9b -
ShouldRespectGracePeriod $false
```

- Para obter detalhes da API, consulte SetInstanceHealthem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`SetInstanceProtection`** com uma CLI

Os exemplos de código a seguir mostram como usar o `SetInstanceProtection`.

CLI

AWS CLI

Exemplo 1: Para alterar a configuração de proteção de instância para uma instância

Este exemplo habilita a proteção de instância na a instância especificada.

```
aws autoscaling set-instance-protection \
    --instance-ids i-061c63c5eb45f0416 \
    --auto-scaling-group-name my-asg --protected-from-scale-in
```

Este comando não produz saída.

Exemplo 2: desativar a configuração de proteção de instância para uma instância

Este exemplo desabilita a proteção de instância na instância especificada.

```
aws autoscaling set-instance-protection \
```

```
    --instance-ids i-061c63c5eb45f0416 \
    --auto-scaling-group-name my-asg \
    --no-protected-from-scale-in
```

Este comando não produz saída.

- Para obter detalhes da API, consulte SetInstanceProtectionem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo habilita a proteção de instância na a instância especificada.

```
Set-ASInstanceProtection -AutoScalingGroupName my-asg -InstanceId i-12345678 -
ProtectedFromScaleIn $true
```

Exemplo 2: Este exemplo desabilita a proteção de instância na instância especificada.

```
Set-ASInstanceProtection -AutoScalingGroupName my-asg -InstanceId i-12345678 -
ProtectedFromScaleIn $false
```

- Para obter detalhes da API, consulte SetInstanceProtectionem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo habilita a proteção de instância na a instância especificada.

```
Set-ASInstanceProtection -AutoScalingGroupName my-asg -InstanceId i-12345678 -
ProtectedFromScaleIn $true
```

Exemplo 2: Este exemplo desabilita a proteção de instância na instância especificada.

```
Set-ASInstanceProtection -AutoScalingGroupName my-asg -InstanceId i-12345678 -
ProtectedFromScaleIn $false
```

- Para obter detalhes da API, consulte SetInstanceProtectionem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Usar **`SuspendProcesses`** com uma CLI

Os exemplos de código a seguir mostram como usar o `SuspendProcesses`.

CLI

AWS CLI

Como suspender os processos do Auto Scaling

Este exemplo suspende o processo de escalabilidade especificado para o grupo do Auto Scaling especificado.

```
aws autoscaling suspend-processes \
    --auto-scaling-group-name my-asg \
    --scaling-processes AlarmNotification
```

Este comando não produz saída.

Para obter mais informações, consulte Suspensão e retomada dos processos de escalabilidade no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte SuspendProcessesem Referência de AWS CLI Comandos.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo suspende o processo de escalabilidade especificado para o grupo do Auto Scaling especificado.

```
Suspend-ASProcess -AutoScalingGroupName my-asg -ScalingProcess
 "AlarmNotification"
```

Exemplo 2: Este exemplo suspende todos os processos do Auto Scaling para o grupo do Auto Scaling especificado.

```
Suspend-ASProcess -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte SuspendProcessesem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo suspende o processo de escalabilidade especificado para o grupo do Auto Scaling especificado.

```
Suspend-ASProcess -AutoScalingGroupName my-asg -ScalingProcess
 "AlarmNotification"
```

Exemplo 2: Este exemplo suspende todos os processos do Auto Scaling para o grupo do Auto Scaling especificado.

```
Suspend-ASProcess -AutoScalingGroupName my-asg
```

- Para obter detalhes da API, consulte SuspendProcessesem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Use **`TerminateInstanceInAutoScalingGroup`** com um AWS SDK ou CLI

Os exemplos de código a seguir mostram como usar o `TerminateInstanceInAutoScalingGroup`.

Exemplos de ações são trechos de código de programas maiores e devem ser executados em contexto. É possível ver essa ação em contexto nos seguintes exemplos de código:

- Conheça os conceitos básicos
- Criar e gerenciar um serviço resiliente

.NET

SDK para .NET (v4)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    /// <summary>
    /// Terminate all instances in the Auto Scaling group in preparation for
    /// deleting the group.
    /// </summary>
    /// <param name="instanceId">The instance Id of the instance to terminate.</
param>
    /// <returns>A Boolean value that indicates the success or failure of
    /// the operation.</returns>
    public async Task<bool> TerminateInstanceInAutoScalingGroupAsync(
        string instanceId)
    {
        var request = new TerminateInstanceInAutoScalingGroupRequest
        {
            InstanceId = instanceId,
            ShouldDecrementDesiredCapacity = false,
        };

        var response = await
 _amazonAutoScaling.TerminateInstanceInAutoScalingGroupAsync(request);

        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
            Console.WriteLine($"You have terminated the instance: {instanceId}");
            return true;
        }

        Console.WriteLine($"Could not terminate {instanceId}");
        return false;
    }
```

- Para obter detalhes da API, consulte TerminateInstanceInAutoScalingGroupa Referência AWS SDK para .NET da API.

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

        Aws::AutoScaling::Model::TerminateInstanceInAutoScalingGroupRequest
request;
        request.SetInstanceId(instanceIDs[instanceNumber - 1]);
        request.SetShouldDecrementDesiredCapacity(false);

        Aws::AutoScaling::Model::TerminateInstanceInAutoScalingGroupOutcome
outcome =
                autoScalingClient.TerminateInstanceInAutoScalingGroup(request);

        if (outcome.IsSuccess()) {
            std::cout << "Waiting for EC2 instance with ID '"
                      << instanceIDs[instanceNumber - 1] << "' to terminate..."
                      << std::endl;
        }
        else {
            std::cerr << "Error with
AutoScaling::TerminateInstanceInAutoScalingGroup. "
                      << outcome.GetError().GetMessage()
                      << std::endl;

        }
```

- Para obter detalhes da API, consulte TerminateInstanceInAutoScalingGroupa Referência AWS SDK para C++ da API.

CLI

AWS CLI

Como encerrar uma instância em um grupo do Auto Scaling

Este exemplo encerra a instância especificada do grupo do Auto Scaling especificado sem atualizar o tamanho do grupo. O Amazon EC2 Auto Scaling executa uma instância de substituição após o encerramento da instância especificada.

```
aws autoscaling terminate-instance-in-auto-scaling-group \
    --instance-id i-061c63c5eb45f0416 \
    --no-should-decrement-desired-capacity
```

Saída:

```
{
    "Activities": [
        {
            "ActivityId": "8c35d601-793c-400c-fcd0-f64a27530df7",
            "AutoScalingGroupName": "my-asg",
            "Description": "Terminating EC2 instance: i-061c63c5eb45f0416",
            "Cause": "",
            "StartTime": "2020-10-31T20:34:25.680Z",
            "StatusCode": "InProgress",
            "Progress": 0,
            "Details": "{\"Subnet ID\":\"subnet-6194ea3b\",\"Availability Zone\":
\"us-west-2c\"}"
        }
    ]
}
```

- Para obter detalhes da API, consulte TerminateInstanceInAutoScalingGroupem Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
   public static void terminateInstanceInAutoScalingGroup(AutoScalingClient
autoScalingClient, String instanceId) {
       try {
           TerminateInstanceInAutoScalingGroupRequest request =
TerminateInstanceInAutoScalingGroupRequest.builder()
                   .instanceId(instanceId)
                   .shouldDecrementDesiredCapacity(false)
                   .build();

           autoScalingClient.terminateInstanceInAutoScalingGroup(request);
           System.out.println("You have terminated instance " + instanceId);

       } catch (AutoScalingException e) {
           System.err.println(e.awsErrorDetails().errorMessage());
           System.exit(1);
       }
   }
```

- Para obter detalhes da API, consulte TerminateInstanceInAutoScalingGroupa Referência AWS SDK for Java 2.x da API.

Kotlin

SDK para Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun terminateInstanceInAutoScalingGroup(instanceIdVal: String) {
    val request =
        TerminateInstanceInAutoScalingGroupRequest {
            instanceId = instanceIdVal
            shouldDecrementDesiredCapacity = false
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.terminateInstanceInAutoScalingGroup(request)
        println("You have terminated instance $instanceIdVal")
    }
}
```

- Para obter detalhes da API, consulte a TerminateInstanceInAutoScalingGroupreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function terminateInstanceInAutoScalingGroup(
        $instanceId,
        $shouldDecrementDesiredCapacity = true,
        $attempts = 0
    ) {
        try {
            return $this->autoScalingClient-
>terminateInstanceInAutoScalingGroup([
                'InstanceId' => $instanceId,
                'ShouldDecrementDesiredCapacity' =>
 $shouldDecrementDesiredCapacity,
            ]);
        } catch (AutoScalingException $exception) {
```

```
            if ($exception->getAwsErrorCode() == "ScalingActivityInProgress" &&
 $attempts < 5) {
                error_log("Cannot terminate an instance while it is still
 pending. Waiting then trying again.");
                sleep(5 * (1 + $attempts));
                return $this->terminateInstanceInAutoScalingGroup(
                    $instanceId,
                    $shouldDecrementDesiredCapacity,
                    ++$attempts
                );
            } else {
                throw $exception;
            }
        }
    }
```

- Para obter detalhes da API, consulte TerminateInstanceInAutoScalingGroupa Referência AWS SDK para PHP da API.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo encerra a instância especificada e diminui a capacidade desejada de seu grupo do Auto Scaling para que o Auto Scaling não inicie uma instância substituta.

```
Stop-ASInstanceInAutoScalingGroup -InstanceId i-93633f9b -
ShouldDecrementDesiredCapacity $true
```

Saída:

```
ActivityId           : 2e40d9bd-1902-444c-abf3-6ea0002efdc5
AutoScalingGroupName :
Cause                : At 2015-11-22T16:09:03Z instance i-93633f9b was taken out
 of service in response to a user
                       request, shrinking the capacity from 2 to 1.
Description          : Terminating EC2 instance: i-93633f9b
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 0
```

```
StartTime            : 11/22/2015 8:09:03 AM
StatusCode           : InProgress
StatusMessage        :
```

Exemplo 2: Este exemplo encerra a instância especificada sem diminuir a capacidade desejada de seu grupo do Auto Scaling. Auto Scaling inicia uma nova instância de substituição.

```
Stop-ASInstanceInAutoScalingGroup -InstanceId i-93633f9b -
ShouldDecrementDesiredCapacity $false
```

Saída:

```
ActivityId           : 2e40d9bd-1902-444c-abf3-6ea0002efdc5
AutoScalingGroupName :
Cause                : At 2015-11-22T16:09:03Z instance i-93633f9b was taken out
 of service in response to a user
                       request.
Description          : Terminating EC2 instance: i-93633f9b
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 0
StartTime            : 11/22/2015 8:09:03 AM
StatusCode           : InProgress
StatusMessage        :
```

- Para obter detalhes da API, consulte TerminateInstanceInAutoScalingGroupem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo encerra a instância especificada e diminui a capacidade desejada de seu grupo do Auto Scaling para que o Auto Scaling não inicie uma instância substituta.

```
Stop-ASInstanceInAutoScalingGroup -InstanceId i-93633f9b -
ShouldDecrementDesiredCapacity $true
```

Saída:

```
ActivityId           : 2e40d9bd-1902-444c-abf3-6ea0002efdc5
```

```
AutoScalingGroupName :
Cause                : At 2015-11-22T16:09:03Z instance i-93633f9b was taken out
 of service in response to a user
                       request, shrinking the capacity from 2 to 1.
Description          : Terminating EC2 instance: i-93633f9b
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 0
StartTime            : 11/22/2015 8:09:03 AM
StatusCode           : InProgress
StatusMessage        :
```

Exemplo 2: Este exemplo encerra a instância especificada sem diminuir a capacidade desejada de seu grupo do Auto Scaling. Auto Scaling inicia uma nova instância de substituição.

```
Stop-ASInstanceInAutoScalingGroup -InstanceId i-93633f9b -
ShouldDecrementDesiredCapacity $false
```

Saída:

```
ActivityId           : 2e40d9bd-1902-444c-abf3-6ea0002efdc5
AutoScalingGroupName :
Cause                : At 2015-11-22T16:09:03Z instance i-93633f9b was taken out
 of service in response to a user
                       request.
Description          : Terminating EC2 instance: i-93633f9b
Details              : {"Availability Zone":"us-west-2b","Subnet
 ID":"subnet-5264e837"}
EndTime              :
Progress             : 0
StartTime            : 11/22/2015 8:09:03 AM
StatusCode           : InProgress
StatusMessage        :
```

- Para obter detalhes da API, consulte TerminateInstanceInAutoScalingGroupem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def terminate_instance(
        self, instance_id: str, decrease_capacity: bool
    ) -> Dict[str, Any]:
        """
        Stops an instance.

        :param instance_id: The ID of the instance to stop.
        :param decrease_capacity: Specifies whether to decrease the desired
 capacity
                                  of the group. When passing True for this
 parameter,
                                  you can stop an instance without having a
 replacement
                                  instance start when the desired capacity
 threshold is
                                  crossed.
        :return: A dictionary containing details of the scaling activity that
 occurs
                 in response to this action.
        :raises ClientError: If there is an error terminating the instance.
        """
        try:
```

```
            response =
self.autoscaling_client.terminate_instance_in_auto_scaling_group(
                InstanceId=instance_id,
ShouldDecrementDesiredCapacity=decrease_capacity
            )
            logger.info(f"Successfully terminated instance {instance_id}.")
            return response["Activity"]

        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            logger.error(f"Failed to terminate instance {instance_id}.")
            if error_code == "ScalingActivityInProgress":
                logger.error(
                    "A scaling activity is currently in progress for the Auto
Scaling group "
                    f"associated with instance '{instance_id}'. "
                    "Please wait for the activity to complete before attempting
to terminate the instance."
                )
            elif error_code == "ResourceInUse":
                logger.error(
                    f"The instance '{instance_id}' or an associated resource is
currently in use "
                    "and cannot be terminated. "
                    "Ensure the instance is not involved in any ongoing processes
and try again."
                )
            logger.error(f"Full error:\n\t{err}")
            raise
```

- Para obter detalhes da API, consulte a TerminateInstanceInAutoScalingGroupReferência da API AWS SDK for Python (Boto3).

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
pub async fn terminate_some_instance(&self) -> Result<(), ScenarioError> {
    // Retrieve a list of instances in the auto scaling group.
    let auto_scaling_group = self.get_group().await?;
    let instances = auto_scaling_group.instances();
    // Or use other logic to find an instance to terminate.
    let instance = instances.first();
    if let Some(instance) = instance {
        let instance_id = if let Some(instance_id) = instance.instance_id() {
            instance_id
        } else {
            return Err(ScenarioError::with("Missing instance id"));
        };
        let termination = self
            .ec2
            .terminate_instances()
            .instance_ids(instance_id)
            .send()
            .await;
        if let Err(err) = termination {
            Err(ScenarioError::new(
                "There was a problem terminating an instance",
                &err,
            ))
        } else {
            Ok(())
        }
    } else {
        Err(ScenarioError::with("There was no instance to terminate"))
    }
}

async fn get_group(&self) -> Result<AutoScalingGroup, ScenarioError> {
```

```
        let describe_auto_scaling_groups = self
            .autoscaling
            .describe_auto_scaling_groups()
            .auto_scaling_group_names(self.auto_scaling_group_name.clone())
            .send()
            .await;

        if let Err(err) = describe_auto_scaling_groups {
            return Err(ScenarioError::new(
                format!(
                    "Failed to get status of autoscaling group {}",
                    self.auto_scaling_group_name.clone()
                )
                .as_str(),
                &err,
            ));
        }

        let describe_auto_scaling_groups_output =
describe_auto_scaling_groups.unwrap();
        let auto_scaling_groups =
describe_auto_scaling_groups_output.auto_scaling_groups();
        let auto_scaling_group = auto_scaling_groups.first();

        if auto_scaling_group.is_none() {
            return Err(ScenarioError::with(format!(
                "Could not find autoscaling group {}",
                self.auto_scaling_group_name.clone()
            )));
        }

        Ok(auto_scaling_group.unwrap().clone())
    }
```

- Para obter detalhes da API, consulte a TerminateInstanceInAutoScalingGroupreferência da API AWS SDK for Rust.

SAP ABAP

SDK para SAP ABAP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
" Example: iv_instance_id = 'i-1234567890abcdef0'
" Example: iv_decrease_capacity = abap_true

TRY.
    DATA(lo_output) = ao_asc->terminateinstinautoscgroup(
      iv_instanceid = iv_instance_id
      iv_shoulddecrementdesiredcap = iv_decrease_capacity ).

    oo_output = lo_output->get_activity( ).

    MESSAGE 'Instance terminated successfully' TYPE 'I'.

  CATCH /aws1/cx_ascscaactivityinprg00 INTO DATA(lo_activity_in_progress).
    RAISE EXCEPTION lo_activity_in_progress.
  CATCH /aws1/cx_ascresrccontionfault INTO DATA(lo_contention).
    RAISE EXCEPTION lo_contention.
  CATCH /aws1/cx_rt_generic INTO DATA(lo_generic_exception).
    RAISE EXCEPTION lo_generic_exception.
ENDTRY.
```

- Para obter detalhes da API, consulte a TerminateInstanceInAutoScalingGroupreferência da API AWS SDK for SAP ABAP.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Use **`UpdateAutoScalingGroup`** com um AWS SDK ou CLI

Os exemplos de código a seguir mostram como usar o `UpdateAutoScalingGroup`.

Exemplos de ações são trechos de código de programas maiores e devem ser executados em contexto. É possível ver essa ação em contexto nos seguintes exemplos de código:

- Conheça os conceitos básicos
- Criar e gerenciar um serviço resiliente

.NET

SDK para .NET (v4)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    /// <summary>
    /// Update the capacity of an Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Auto Scaling group.</param>
    /// <param name="launchTemplateName">The name of the EC2 launch template.</
param>
    /// <param name="maxSize">The maximum number of instances that can be
    /// created for the Auto Scaling group.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> UpdateAutoScalingGroupAsync(
        string groupName,
        string launchTemplateName,
        int maxSize)
    {
        var templateSpecification = new LaunchTemplateSpecification
        {
            LaunchTemplateName = launchTemplateName,
        };

        var groupRequest = new UpdateAutoScalingGroupRequest
        {
            MaxSize = maxSize,
            AutoScalingGroupName = groupName,
            LaunchTemplate = templateSpecification,
        };
```

```
        var response = await
_amazonAutoScaling.UpdateAutoScalingGroupAsync(groupRequest);
        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
            Console.WriteLine($"You successfully updated the Auto Scaling group
{groupName}.");
            return true;
        }
        else
        {
            return false;
        }
    }
```

- Para obter detalhes da API, consulte UpdateAutoScalingGroupa Referência AWS SDK para .NET da API.

C++

SDK para C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

        Aws::AutoScaling::Model::UpdateAutoScalingGroupRequest request;
        request.SetAutoScalingGroupName(groupName);
        request.SetMaxSize(3);

        Aws::AutoScaling::Model::UpdateAutoScalingGroupOutcome outcome =
```

```
                autoScalingClient.UpdateAutoScalingGroup(request);

        if (!outcome.IsSuccess()) {
            std::cerr << "Error with AutoScaling::UpdateAutoScalingGroup. "
                      << outcome.GetError().GetMessage()
                      << std::endl;

        }
```

- Para obter detalhes da API, consulte UpdateAutoScalingGroupa Referência AWS SDK para C++ da API.

CLI

AWS CLI

Exemplo 1: como atualizar os limites de tamanho de um grupo do Auto Scaling

Este exemplo atualiza o grupo do Auto Scaling especificado com um tamanho mínimo de 2 e máximo de 10.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --min-size 2 \
    --max-size 10
```

Este comando não produz saída.

Para obter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: como adicionar verificações de integridade do Elastic Load Balancing e especificar quais zonas de disponibilidade e sub-redes usar

Este exemplo atualiza o grupo do Auto Scaling especificado para adicionar verificações de integridade do Elastic Load Balancing. Esse comando também atualiza o valor de `--vpc-zone-identifier` com uma lista de sub-redes IDs em várias zonas de disponibilidade.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --health-check-type ELB \
```

```
    --health-check-grace-period 600 \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Este comando não produz saída.

Para obter mais informações, consulte Elastic Load Balancing e Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 3: como atualizar o grupo de posicionamento e a política de encerramento

Este exemplo atualiza o grupo de posicionamento e a política de encerramento que devem ser usados.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --placement-group my-placement-group \
    --termination-policies "OldestInstance"
```

Este comando não produz saída.

Para obter mais informações, consulte Grupos do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 4: como usar a versão mais recente do modelo de execução

Este exemplo atualiza o grupo do Auto Scaling especificado para que use a versão mais recente do modelo de execução especificado.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12,Version='$Latest'
```

Este comando não produz saída.

Para obter mais informações, consulte Modelos de execução no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 5: como usar uma versão específica do modelo de execução

Este exemplo atualiza o grupo do Auto Scaling especificado para que use uma versão específica do modelo de execução em vez da versão mais recente ou padrão.

```
aws autoscaling update-auto-scaling-group \
```

```
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateName=my-template-for-auto-scaling,Version='2'
```

Este comando não produz saída.

Para obter mais informações, consulte Modelos de execução no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 6: como definir uma política de instâncias mistas e habilitar o rebalanceamento de capacidade

Este exemplo atualiza o grupo do Auto Scaling especificado para que use uma política de instâncias mistas e permita o rebalanceamento de capacidade. Essa estrutura permite especificar grupos com capacidades spot e sob demanda e usar modelos de execução diferentes para arquiteturas diferentes.

```
aws autoscaling update-auto-scaling-group \
    --cli-input-json file://~/config.json
```

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "CapacityRebalance": true,
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template-for-x86",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "c6g.large",
                    "LaunchTemplateSpecification": {
                        "LaunchTemplateName": "my-launch-template-for-arm",
                        "Version": "$Latest"
                    }
                },
                {
                    "InstanceType": "c5.large"
                },
                {
```

```
                    "InstanceType": "c5a.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 50,
            "SpotAllocationStrategy": "capacity-optimized"
        }
    }
}
```

Este comando não produz saída.

Para obter mais informações, consulte Grupos de Auto Scaling com vários tipos de instância e opções de compra no Manual do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte UpdateAutoScalingGroupem Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
   public static void updateAutoScalingGroup(AutoScalingClient
autoScalingClient, String groupName,
           String launchTemplateName) {
       try {
           AutoScalingWaiter waiter = autoScalingClient.waiter();
           LaunchTemplateSpecification templateSpecification =
LaunchTemplateSpecification.builder()
                   .launchTemplateName(launchTemplateName)
                   .build();

           UpdateAutoScalingGroupRequest groupRequest =
UpdateAutoScalingGroupRequest.builder()
                   .maxSize(3)
```

```
                    .autoScalingGroupName(groupName)
                    .launchTemplate(templateSpecification)
                    .build();

            autoScalingClient.updateAutoScalingGroup(groupRequest);
            DescribeAutoScalingGroupsRequest groupsRequest =
DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();

            WaiterResponse<DescribeAutoScalingGroupsResponse> waiterResponse =
waiter
                    .waitUntilGroupInService(groupsRequest);
            waiterResponse.matched().response().ifPresent(System.out::println);
            System.out.println("You successfully updated the auto scaling group
" + groupName);

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
```

- Para obter detalhes da API, consulte UpdateAutoScalingGroupa Referência AWS SDK for Java 2.x da API.

Kotlin

SDK para Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun updateAutoScalingGroup(
    groupName: String,
    launchTemplateNameVal: String,
    serviceLinkedRoleARNVal: String,
```

```
) {
    val templateSpecification =
        LaunchTemplateSpecification {
            launchTemplateName = launchTemplateNameVal
        }

    val groupRequest =
        UpdateAutoScalingGroupRequest {
            maxSize = 3
            serviceLinkedRoleArn = serviceLinkedRoleARNVal
            autoScalingGroupName = groupName
            launchTemplate = templateSpecification
        }

    val groupsRequestWaiter =
        DescribeAutoScalingGroupsRequest {
            autoScalingGroupNames = listOf(groupName)
        }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.updateAutoScalingGroup(groupRequest)
        autoScalingClient.waitUntilGroupExists(groupsRequestWaiter)
        println("You successfully updated the Auto Scaling group  $groupName")
    }
}
```

- Para obter detalhes da API, consulte a UpdateAutoScalingGroupreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function updateAutoScalingGroup($autoScalingGroupName, $args)
```

```
    {
        if (array_key_exists('MaxSize', $args)) {
            $maxSize = ['MaxSize' => $args['MaxSize']];
        } else {
            $maxSize = [];
        }
        if (array_key_exists('MinSize', $args)) {
            $minSize = ['MinSize' => $args['MinSize']];
        } else {
            $minSize = [];
        }
        $parameters = ['AutoScalingGroupName' => $autoScalingGroupName];
        $parameters = array_merge($parameters, $minSize, $maxSize);
        return $this->autoScalingClient->updateAutoScalingGroup($parameters);
    }
```

- Para obter detalhes da API, consulte UpdateAutoScalingGroupa Referência AWS SDK para PHP da API.

PowerShell

Ferramentas para PowerShell V4

Exemplo 1: Este exemplo atualiza o tamanho mínimo e máximo do grupo do Auto Scaling especificado.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -MaxSize 5 -MinSize 1
```

Exemplo 2: Este exemplo atualiza o período de espera padrão do grupo do Auto Scaling especificado.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -DefaultCooldown 10
```

Exemplo 3: Este exemplo atualiza as zonas de disponibilidade do grupo do Auto Scaling especificado.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -AvailabilityZone @("us-
west-2a", "us-west-2b")
```

Exemplo 4: Este exemplo atualiza o grupo do Auto Scaling especificado para usar verificações de integridade do Elastic Load Balancing.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -HealthCheckType ELB -
HealthCheckGracePeriod 60
```

- Para obter detalhes da API, consulte UpdateAutoScalingGroupem Referência de Ferramentas da AWS para PowerShell cmdlet (V4).

Ferramentas para PowerShell V5

Exemplo 1: Este exemplo atualiza o tamanho mínimo e máximo do grupo do Auto Scaling especificado.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -MaxSize 5 -MinSize 1
```

Exemplo 2: Este exemplo atualiza o período de espera padrão do grupo do Auto Scaling especificado.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -DefaultCooldown 10
```

Exemplo 3: Este exemplo atualiza as zonas de disponibilidade do grupo do Auto Scaling especificado.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -AvailabilityZone @("us-
west-2a", "us-west-2b")
```

Exemplo 4: Este exemplo atualiza o grupo do Auto Scaling especificado para usar verificações de integridade do Elastic Load Balancing.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -HealthCheckType ELB -
HealthCheckGracePeriod 60
```

- Para obter detalhes da API, consulte UpdateAutoScalingGroupem Referência de Ferramentas da AWS para PowerShell cmdlet (V5).

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def update_group(self, group_name: str, **kwargs: Any) -> None:
        """
        Updates an Auto Scaling group.

        :param group_name: The name of the group to update.
        :param kwargs: Keyword arguments to pass through to the service.
        :return: None
        :raises ClientError: If there is an error updating the Auto Scaling
 group.
        """
        try:
            self.autoscaling_client.update_auto_scaling_group(
                AutoScalingGroupName=group_name, **kwargs
            )
            logger.info(f"Successfully updated Auto Scaling group {group_name}.")

        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            logger.error(f"Failed to update Auto Scaling group {group_name}.")
            if error_code == "ResourceInUse":
                logger.error(
```

```
                    "The Auto Scaling group '%s' is currently in use and cannot
 be modified. Please try again later.",
                    group_name,
                )
            elif error_code == "ScalingActivityInProgress":
                logger.error(
                    f"A scaling activity is currently in progress for the Auto
 Scaling group '{group_name}'."
                    "Please wait for the activity to complete before attempting
 to update the group."
                )
            logger.error(f"Full error:\n\t{err}")
            raise
```

- Para obter detalhes da API, consulte a UpdateAutoScalingGroupReferência da API AWS SDK for Python (Boto3).

Rust

SDK para Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
async fn update_group(client: &Client, name: &str, size: i32) -> Result<(),
 Error> {
    client
        .update_auto_scaling_group()
        .auto_scaling_group_name(name)
        .max_size(size)
        .send()
        .await?;

    println!("Updated AutoScaling group");

    Ok(())
```

```
}
```

- Para obter detalhes da API, consulte a UpdateAutoScalingGroupreferência da API AWS SDK for Rust.

SAP ABAP

SDK para SAP ABAP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
" Example: iv_group_name = 'my-auto-scaling-group'
" Example: iv_max_size = 5

TRY.
    ao_asc->updateautoscalinggroup(
      iv_autoscalinggroupname = iv_group_name
      iv_maxsize = iv_max_size
      iv_minsize = iv_min_size ).

    MESSAGE 'Auto Scaling group updated successfully' TYPE 'I'.

  CATCH /aws1/cx_ascresrccontionfault INTO DATA(lo_contention).
    RAISE EXCEPTION lo_contention.
  CATCH /aws1/cx_ascscaactivityinprg00 INTO DATA(lo_activity_in_progress).
    RAISE EXCEPTION lo_activity_in_progress.
  CATCH /aws1/cx_rt_generic INTO DATA(lo_generic_exception).
    RAISE EXCEPTION lo_generic_exception.
ENDTRY.
```

- Para obter detalhes da API, consulte a UpdateAutoScalingGroupreferência da API AWS SDK for SAP ABAP.

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Cenários para o Auto Scaling usando AWS SDKs

Os exemplos de código a seguir mostram como implementar cenários comuns no Auto Scaling com. AWS SDKs Esses cenários mostram como realizar tarefas específicas chamando várias funções no Aurora ou combinadas com outros Serviços da AWS. Cada cenário inclui um link para o código-fonte completo, onde podem ser encontradas instruções sobre como configurar e executar o código.

Os cenários têm como alvo um nível intermediário de experiência para ajudar você a compreender ações de serviço em contexto.

Exemplos

- Crie e gerencie um serviço resiliente usando um SDK AWS

Crie e gerencie um serviço resiliente usando um SDK AWS

Os exemplos de código a seguir mostram como criar um serviço da Web com carga balanceada que retorna recomendações de livros, filmes e músicas. O exemplo mostra como o serviço responde a falhas e como é possível reestruturá-lo para gerar mais resiliência em caso de falhas.

- Use um grupo do Amazon EC2 Auto Scaling para criar instâncias do Amazon Elastic Compute Cloud (Amazon EC2) com base em um modelo de execução e para manter o número de instâncias em um intervalo especificado.
- Gerencie e distribua solicitações HTTP com o Elastic Load Balancing.
- Monitore a integridade das instâncias em um grupo do Auto Scaling e encaminhe solicitações somente para instâncias íntegras.
- Execute um servidor Web Python em cada instância do EC2 para lidar com solicitações HTTP. O servidor Web responde com recomendações e verificações de integridade.
- Simule um serviço de recomendação com uma tabela do Amazon DynamoDB.
- Controle a resposta do servidor web às solicitações e verificações de saúde atualizando AWS Systems Manager os parâmetros.

.NET

SDK para .NET

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

Execute o cenário interativo em um prompt de comando.

```
    static async Task Main(string[] args)
    {
        _configuration = new ConfigurationBuilder()
            .SetBasePath(Directory.GetCurrentDirectory())
            .AddJsonFile("settings.json") // Load settings from .json file.
            .AddJsonFile("settings.local.json",
                true) // Optionally, load local settings.
            .Build();


        // Set up dependency injection for the AWS services.
        using var host = Host.CreateDefaultBuilder(args)
            .ConfigureLogging(logging =>
                logging.AddFilter("System", LogLevel.Debug)
                    .AddFilter<DebugLoggerProvider>("Microsoft",
LogLevel.Information)
                    .AddFilter<ConsoleLoggerProvider>("Microsoft",
LogLevel.Trace))
            .ConfigureServices((_, services) =>
                services.AddAWSService<IAmazonIdentityManagementService>()
                    .AddAWSService<IAmazonDynamoDB>()
                    .AddAWSService<IAmazonElasticLoadBalancingV2>()
                    .AddAWSService<IAmazonSimpleSystemsManagement>()
                    .AddAWSService<IAmazonAutoScaling>()
                    .AddAWSService<IAmazonEC2>()
                    .AddTransient<AutoScalerWrapper>()
                    .AddTransient<ElasticLoadBalancerWrapper>()
                    .AddTransient<SmParameterWrapper>()
                    .AddTransient<Recommendations>()
                    .AddSingleton<IConfiguration>(_configuration)
            )
```

```
            .Build();

        ServicesSetup(host);
        ResourcesSetup();

        try
        {
            Console.WriteLine(new string('-', 80));
            Console.WriteLine("Welcome to the Resilient Architecture Example
Scenario.");
            Console.WriteLine(new string('-', 80));
            await Deploy(true);

            Console.WriteLine("Now let's begin the scenario.");
            Console.WriteLine(new string('-', 80));
            await Demo(true);

            Console.WriteLine(new string('-', 80));
            Console.WriteLine("Finally, let's clean up our resources.");
            Console.WriteLine(new string('-', 80));

            await DestroyResources(true);

            Console.WriteLine(new string('-', 80));
            Console.WriteLine("Resilient Architecture Example Scenario is
complete.");
            Console.WriteLine(new string('-', 80));
        }
        catch (Exception ex)
        {
            Console.WriteLine(new string('-', 80));
            Console.WriteLine($"There was a problem running the scenario:
{ex.Message}");
            await DestroyResources(true);
            Console.WriteLine(new string('-', 80));
        }
    }

    /// <summary>
    /// Setup any common resources, also used for integration testing.
    /// </summary>
    public static void ResourcesSetup()
    {
        _httpClient = new HttpClient();
```

```
    }

    /// <summary>
    /// Populate the services for use within the console application.
    /// </summary>
    /// <param name="host">The services host.</param>
    private static void ServicesSetup(IHost host)
    {
        _elasticLoadBalancerWrapper =
host.Services.GetRequiredService<ElasticLoadBalancerWrapper>();
        _iamClient =
host.Services.GetRequiredService<IAmazonIdentityManagementService>();
        _recommendations = host.Services.GetRequiredService<Recommendations>();
        _autoScalerWrapper =
host.Services.GetRequiredService<AutoScalerWrapper>();
        _smParameterWrapper =
host.Services.GetRequiredService<SmParameterWrapper>();
    }

    /// <summary>
    /// Deploy necessary resources for the scenario.
    /// </summary>
    /// <param name="interactive">True to run as interactive.</param>
    /// <returns>True if successful.</returns>
    public static async Task<bool> Deploy(bool interactive)
    {
        var protocol = "HTTP";
        var port = 80;
        var sshPort = 22;

        Console.WriteLine(
            "\nFor this demo, we'll use the AWS SDK for .NET to create several
AWS resources\n" +
            "to set up a load-balanced web service endpoint and explore some ways
to make it resilient\n" +
            "against various kinds of failures.\n\n" +
            "Some of the resources create by this demo are:\n");

        Console.WriteLine(
            "\t* A DynamoDB table that the web service depends on to provide
book, movie, and song recommendations.");
        Console.WriteLine(
            "\t* An EC2 launch template that defines EC2 instances that each
contain a Python web server.");
```

```
        Console.WriteLine(
            "\t* An EC2 Auto Scaling group that manages EC2 instances across
several Availability Zones.");
        Console.WriteLine(
            "\t* An Elastic Load Balancing (ELB) load balancer that targets the
Auto Scaling group to distribute requests.");
        Console.WriteLine(new string('-', 80));
        Console.WriteLine("Press Enter when you're ready to start deploying
resources.");
        if (interactive)
            Console.ReadLine();

        // Create and populate the DynamoDB table.
        var databaseTableName = _configuration["databaseName"];
        var recommendationsPath = Path.Join(_configuration["resourcePath"],
            "recommendations_objects.json");
        Console.WriteLine($"Creating and populating a DynamoDB table named
{databaseTableName}.");
        await _recommendations.CreateDatabaseWithName(databaseTableName);
        await _recommendations.PopulateDatabase(databaseTableName,
recommendationsPath);
        Console.WriteLine(new string('-', 80));

        // Create the EC2 Launch Template.

        Console.WriteLine(
            $"Creating an EC2 launch template that runs
'server_startup_script.sh' when an instance starts.\n"
            + "\nThis script starts a Python web server defined in the
`server.py` script. The web server\n"
            + "listens to HTTP requests on port 80 and responds to requests to
'/' and to '/healthcheck'.\n"
            + "For demo purposes, this server is run as the root user. In
production, the best practice is to\n"
            + "run a web server, such as Apache, with least-privileged
credentials.");
        Console.WriteLine(
            "\nThe template also defines an IAM policy that each instance uses to
assume a role that grants\n"
            + "permissions to access the DynamoDB recommendation table and
Systems Manager parameters\n"
            + "that control the flow of the demo.");

        var startupScriptPath = Path.Join(_configuration["resourcePath"],
```

```
            "server_startup_script.sh");
        var instancePolicyPath = Path.Join(_configuration["resourcePath"],
            "instance_policy.json");
        await _autoScalerWrapper.CreateTemplate(startupScriptPath,
instancePolicyPath);
        Console.WriteLine(new string('-', 80));

        Console.WriteLine(
            "Creating an EC2 Auto Scaling group that maintains three EC2
instances, each in a different\n"
            + "Availability Zone.\n");
        var zones = await _autoScalerWrapper.DescribeAvailabilityZones();
        await _autoScalerWrapper.CreateGroupOfSize(3,
_autoScalerWrapper.GroupName, zones);
        Console.WriteLine(new string('-', 80));

        Console.WriteLine(
            "At this point, you have EC2 instances created. Once each instance
starts, it listens for\n"
            + "HTTP requests. You can see these instances in the console or
continue with the demo.\n");

        Console.WriteLine(new string('-', 80));
        Console.WriteLine("Press Enter when you're ready to continue.");
        if (interactive)
            Console.ReadLine();

        Console.WriteLine("Creating variables that control the flow of the
demo.");
        await _smParameterWrapper.Reset();

        Console.WriteLine(
            "\nCreating an Elastic Load Balancing target group and load balancer.
The target group\n"
            + "defines how the load balancer connects to instances. The load
balancer provides a\n"
            + "single endpoint where clients connect and dispatches requests to
instances in the group.");

        var defaultVpc = await _autoScalerWrapper.GetDefaultVpc();
        var subnets = await
_autoScalerWrapper.GetAllVpcSubnetsForZones(defaultVpc.VpcId, zones);
        var subnetIds = subnets.Select(s => s.SubnetId).ToList();
```

```
        var targetGroup = await
 _elasticLoadBalancerWrapper.CreateTargetGroupOnVpc(_elasticLoadBalancerWrapper.TargetGro
 protocol, port, defaultVpc.VpcId);

        await
 _elasticLoadBalancerWrapper.CreateLoadBalancerAndListener(_elasticLoadBalancerWrapper.Lo
 subnetIds, targetGroup);
        await
 _autoScalerWrapper.AttachLoadBalancerToGroup(_autoScalerWrapper.GroupName,
 targetGroup.TargetGroupArn);
        Console.WriteLine("\nVerifying access to the load balancer endpoint...");
        var endPoint = await
 _elasticLoadBalancerWrapper.GetEndpointForLoadBalancerByName(_elasticLoadBalancerWrapper
        var loadBalancerAccess = await
 _elasticLoadBalancerWrapper.VerifyLoadBalancerEndpoint(endPoint);

        if (!loadBalancerAccess)
        {
            Console.WriteLine("\nCouldn't connect to the load balancer, verifying
 that the port is open...");

            var ipString = await _httpClient.GetStringAsync("https://
checkip.amazonaws.com");
            ipString = ipString.Trim();

            var defaultSecurityGroup = await
 _autoScalerWrapper.GetDefaultSecurityGroupForVpc(defaultVpc);
            var portIsOpen =
 _autoScalerWrapper.VerifyInboundPortForGroup(defaultSecurityGroup, port,
 ipString);
            var sshPortIsOpen =
 _autoScalerWrapper.VerifyInboundPortForGroup(defaultSecurityGroup, sshPort,
 ipString);

            if (!portIsOpen)
            {
                Console.WriteLine(
                    "\nFor this example to work, the default security group for
 your default VPC must\n"
                    + "allows access from this computer. You can either add it
 automatically from this\n"
                    + "example or add it yourself using the AWS Management
 Console.\n");
```

```
                if (!interactive || GetYesNoResponse(
                        "Do you want to add a rule to the security group to allow
inbound traffic from your computer's IP address?"))
                {
                    await
_autoScalerWrapper.OpenInboundPort(defaultSecurityGroup.GroupId, port,
ipString);
                }
            }

            if (!sshPortIsOpen)
            {
                if (!interactive || GetYesNoResponse(
                        "Do you want to add a rule to the security group to allow
inbound SSH traffic for debugging from your computer's IP address?"))
                {
                    await
_autoScalerWrapper.OpenInboundPort(defaultSecurityGroup.GroupId, sshPort,
ipString);
                }
            }
            loadBalancerAccess = await
_elasticLoadBalancerWrapper.VerifyLoadBalancerEndpoint(endPoint);
        }

        if (loadBalancerAccess)
        {
            Console.WriteLine("Your load balancer is ready. You can access it by
browsing to:");
            Console.WriteLine($"\thttp://{endPoint}\n");
        }
        else
        {
            Console.WriteLine(
                "\nCouldn't get a successful response from the load balancer
endpoint. Troubleshoot by\n"
                + "manually verifying that your VPC and security group are
configured correctly and that\n"
                + "you can successfully make a GET request to the load balancer
endpoint:\n");
            Console.WriteLine($"\thttp://{endPoint}\n");
        }
        Console.WriteLine(new string('-', 80));
```

```
        Console.WriteLine("Press Enter when you're ready to continue with the
demo.");
        if (interactive)
            Console.ReadLine();
        return true;
    }

    /// <summary>
    /// Demonstrate the steps of the scenario.
    /// </summary>
    /// <param name="interactive">True to run as an interactive scenario.</param>
    /// <returns>Async task.</returns>
    public static async Task<bool> Demo(bool interactive)
    {
        var ssmOnlyPolicy = Path.Join(_configuration["resourcePath"],
            "ssm_only_policy.json");

        Console.WriteLine(new string('-', 80));
        Console.WriteLine("Resetting parameters to starting values for demo.");
        await _smParameterWrapper.Reset();

        Console.WriteLine("\nThis part of the demonstration shows how to toggle
different parts of the system\n" +
                          "to create situations where the web service fails, and
shows how using a resilient\n" +
                          "architecture can keep the web service running in spite
of these failures.");
        Console.WriteLine(new string('-', 88));
        Console.WriteLine("At the start, the load balancer endpoint returns
recommendations and reports that all targets are healthy.");
        if (interactive)
            await DemoActionChoices();

        Console.WriteLine($"The web service running on the EC2 instances gets
recommendations by querying a DynamoDB table.\n" +
                          $"The table name is contained in a Systems Manager
parameter named '{_smParameterWrapper.TableParameter}'.\n" +
                          $"To simulate a failure of the recommendation service,
let's set this parameter to name a non-existent table.\n");
        await
_smParameterWrapper.PutParameterByName(_smParameterWrapper.TableParameter,
"this-is-not-a-table");
        Console.WriteLine("\nNow, sending a GET request to the load balancer
endpoint returns a failure code. But, the service reports as\n" +
```

```
                            "healthy to the load balancer because shallow health
checks don't check for failure of the recommendation service.");
        if (interactive)
            await DemoActionChoices();

        Console.WriteLine("Instead of failing when the recommendation service
fails, the web service can return a static response.");
        Console.WriteLine("While this is not a perfect solution, it presents the
customer with a somewhat better experience than failure.");

        await
_smParameterWrapper.PutParameterByName(_smParameterWrapper.FailureResponseParameter,
"static");

        Console.WriteLine("\nNow, sending a GET request to the load balancer
endpoint returns a static response.");
        Console.WriteLine("The service still reports as healthy because health
checks are still shallow.");
        if (interactive)
            await DemoActionChoices();

        Console.WriteLine("Let's reinstate the recommendation service.\n");
        await
_smParameterWrapper.PutParameterByName(_smParameterWrapper.TableParameter,
_smParameterWrapper.TableName);
        Console.WriteLine(
            "\nLet's also substitute bad credentials for one of the instances in
the target group so that it can't\n" +
            "access the DynamoDB recommendation table.\n"
        );
        await _autoScalerWrapper.CreateInstanceProfileWithName(
            _autoScalerWrapper.BadCredsPolicyName,
            _autoScalerWrapper.BadCredsRoleName,
            _autoScalerWrapper.BadCredsProfileName,
            ssmOnlyPolicy,
            new List<string> { "AmazonSSMManagedInstanceCore" }
        );
        var instances = await
_autoScalerWrapper.GetInstancesByGroupName(_autoScalerWrapper.GroupName);
        var badInstanceId = instances.First();
        var instanceProfile = await
_autoScalerWrapper.GetInstanceProfile(badInstanceId);
        Console.WriteLine(
```

```
            $"Replacing the profile for instance {badInstanceId} with a profile
that contains\n" +
            "bad credentials...\n"
        );
        await _autoScalerWrapper.ReplaceInstanceProfile(
            badInstanceId,
            _autoScalerWrapper.BadCredsProfileName,
            instanceProfile.AssociationId
        );
        Console.WriteLine(
            "Now, sending a GET request to the load balancer endpoint returns
either a recommendation or a static response,\n" +
            "depending on which instance is selected by the load balancer.\n"
        );
        if (interactive)
            await DemoActionChoices();

        Console.WriteLine("\nLet's implement a deep health check. For this demo,
a deep health check tests whether");
        Console.WriteLine("the web service can access the DynamoDB table that it
depends on for recommendations. Note that");
        Console.WriteLine("the deep health check is only for ELB routing and not
for Auto Scaling instance health.");
        Console.WriteLine("This kind of deep health check is not recommended for
Auto Scaling instance health, because it");
        Console.WriteLine("risks accidental termination of all instances in the
Auto Scaling group when a dependent service fails.");

        Console.WriteLine("\nBy implementing deep health checks, the load
balancer can detect when one of the instances is failing");
        Console.WriteLine("and take that instance out of rotation.");

        await
_smParameterWrapper.PutParameterByName(_smParameterWrapper.HealthCheckParameter,
"deep");

        Console.WriteLine($"\nNow, checking target health indicates that the
instance with bad credentials ({badInstanceId})");
        Console.WriteLine("is unhealthy. Note that it might take a minute or two
for the load balancer to detect the unhealthy");
        Console.WriteLine("instance. Sending a GET request to the load balancer
endpoint always returns a recommendation, because");
        Console.WriteLine("the load balancer takes unhealthy instances out of its
rotation.");
```

```
        if (interactive)
            await DemoActionChoices();

        Console.WriteLine("\nBecause the instances in this demo are controlled by
an auto scaler, the simplest way to fix an unhealthy");
        Console.WriteLine("instance is to terminate it and let the auto scaler
start a new instance to replace it.");

        await _autoScalerWrapper.TryTerminateInstanceById(badInstanceId);

        Console.WriteLine($"\nEven while the instance is terminating and the new
instance is starting, sending a GET");
        Console.WriteLine("request to the web service continues to get a
successful recommendation response because");
        Console.WriteLine("starts and reports as healthy, it is included in the
load balancing rotation.");
        Console.WriteLine("Note that terminating and replacing an instance
typically takes several minutes, during which time you");
        Console.WriteLine("can see the changing health check status until the new
instance is running and healthy.");

        if (interactive)
            await DemoActionChoices();

        Console.WriteLine("\nIf the recommendation service fails now, deep health
checks mean all instances report as unhealthy.");

        await
_smParameterWrapper.PutParameterByName(_smParameterWrapper.TableParameter,
"this-is-not-a-table");

        Console.WriteLine($"\nWhen all instances are unhealthy, the load balancer
continues to route requests even to");
        Console.WriteLine("unhealthy instances, allowing them to fail open and
return a static response rather than fail");
        Console.WriteLine("closed and report failure to the customer.");

        if (interactive)
            await DemoActionChoices();
        await _smParameterWrapper.Reset();

        Console.WriteLine(new string('-', 80));
        return true;
```

```
    }

    /// <summary>
    /// Clean up the resources from the scenario.
    /// </summary>
    /// <param name="interactive">True to ask the user for cleanup.</param>
    /// <returns>Async task.</returns>
    public static async Task<bool> DestroyResources(bool interactive)
    {
        Console.WriteLine(new string('-', 80));
        Console.WriteLine(
            "To keep things tidy and to avoid unwanted charges on your account,
we can clean up all AWS resources\n" +
            "that were created for this demo."
        );

        if (!interactive || GetYesNoResponse("Do you want to clean up all demo
resources? (y/n) "))
        {
            await
_elasticLoadBalancerWrapper.DeleteLoadBalancerByName(_elasticLoadBalancerWrapper.LoadBal
            await
_elasticLoadBalancerWrapper.DeleteTargetGroupByName(_elasticLoadBalancerWrapper.TargetGr
            await
_autoScalerWrapper.TerminateAndDeleteAutoScalingGroupWithName(_autoScalerWrapper.GroupNa
            await
_autoScalerWrapper.DeleteKeyPairByName(_autoScalerWrapper.KeyPairName);
            await
_autoScalerWrapper.DeleteTemplateByName(_autoScalerWrapper.LaunchTemplateName);
            await _autoScalerWrapper.DeleteInstanceProfile(
                _autoScalerWrapper.BadCredsProfileName,
                _autoScalerWrapper.BadCredsRoleName
            );
            await
_recommendations.DestroyDatabaseByName(_recommendations.TableName);
        }
        else
        {
            Console.WriteLine(
                "Ok, we'll leave the resources intact.\n" +
                "Don't forget to delete them when you're done with them or you
might incur unexpected charges."
            );
        }
```

```
        Console.WriteLine(new string('-', 80));
        return true;
    }
```

Crie uma classe que envolva ações do Auto Scaling e do Amazon EC2.

```
/// <summary>
/// Encapsulates Amazon EC2 Auto Scaling and EC2 management methods.
/// </summary>
public class AutoScalerWrapper
{
    private readonly IAmazonAutoScaling _amazonAutoScaling;
    private readonly IAmazonEC2 _amazonEc2;
    private readonly IAmazonSimpleSystemsManagement _amazonSsm;
    private readonly IAmazonIdentityManagementService _amazonIam;
    private readonly ILogger<AutoScalerWrapper> _logger;

    private readonly string _instanceType = "";
    private readonly string _amiParam = "";
    private readonly string _launchTemplateName = "";
    private readonly string _groupName = "";
    private readonly string _instancePolicyName = "";
    private readonly string _instanceRoleName = "";
    private readonly string _instanceProfileName = "";
    private readonly string _badCredsProfileName = "";
    private readonly string _badCredsRoleName = "";
    private readonly string _badCredsPolicyName = "";
    private readonly string _keyPairName = "";

    public string GroupName => _groupName;
    public string KeyPairName => _keyPairName;
    public string LaunchTemplateName => _launchTemplateName;
    public string InstancePolicyName => _instancePolicyName;
    public string BadCredsProfileName => _badCredsProfileName;
    public string BadCredsRoleName => _badCredsRoleName;
    public string BadCredsPolicyName => _badCredsPolicyName;

    /// <summary>
    /// Constructor for the AutoScalerWrapper.
    /// </summary>
    /// <param name="amazonAutoScaling">The injected AutoScaling client.</param>
```

```
    /// <param name="amazonEc2">The injected EC2 client.</param>
    /// <param name="amazonIam">The injected IAM client.</param>
    /// <param name="amazonSsm">The injected SSM client.</param>
    public AutoScalerWrapper(
        IAmazonAutoScaling amazonAutoScaling,
        IAmazonEC2 amazonEc2,
        IAmazonSimpleSystemsManagement amazonSsm,
        IAmazonIdentityManagementService amazonIam,
        IConfiguration configuration,
        ILogger<AutoScalerWrapper> logger)
    {
        _amazonAutoScaling = amazonAutoScaling;
        _amazonEc2 = amazonEc2;
        _amazonSsm = amazonSsm;
        _amazonIam = amazonIam;
        _logger = logger;

        var prefix = configuration["resourcePrefix"];
        _instanceType = configuration["instanceType"];
        _amiParam = configuration["amiParam"];

        _launchTemplateName = prefix + "-template";
        _groupName = prefix + "-group";
        _instancePolicyName = prefix + "-pol";
        _instanceRoleName = prefix + "-role";
        _instanceProfileName = prefix + "-prof";
        _badCredsPolicyName = prefix + "-bc-pol";
        _badCredsRoleName = prefix + "-bc-role";
        _badCredsProfileName = prefix + "-bc-prof";
        _keyPairName = prefix + "-key-pair";
    }

    /// <summary>
    /// Create a policy, role, and profile that is associated with instances with
a specified name.
    /// An instance's associated profile defines a role that is assumed by the
    /// instance.The role has attached policies that specify the AWS permissions
granted to
    /// clients that run on the instance.
    /// </summary>
    /// <param name="policyName">Name to use for the policy.</param>
    /// <param name="roleName">Name to use for the role.</param>
    /// <param name="profileName">Name to use for the profile.</param>
    /// <param name="ssmOnlyPolicyFile">Path to a policy file for SSM.</param>
```

```
    /// <param name="awsManagedPolicies">AWS Managed policies to be attached to
the role.</param>
    /// <returns>The Arn of the profile.</returns>
    public async Task<string> CreateInstanceProfileWithName(
        string policyName,
        string roleName,
        string profileName,
        string ssmOnlyPolicyFile,
        List<string>? awsManagedPolicies = null)
    {

        var assumeRoleDoc = "{" +
                                    "\"Version\": \"2012-10-17\"," +
                                    "\"Statement\": [{" +
                                         "\"Effect\": \"Allow\"," +
                                         "\"Principal\": {" +
                                         "\"Service\": [" +
                                             "\"ec2.amazonaws.com\"" +
                                         "]" +
                                         "}," +
                                    "\"Action\": \"sts:AssumeRole\"" +
                                    "}]" +
                                "}";

        var policyDocument = await File.ReadAllTextAsync(ssmOnlyPolicyFile);

        var policyArn = "";

        try
        {
            var createPolicyResult = await _amazonIam.CreatePolicyAsync(
                new CreatePolicyRequest
                {
                    PolicyName = policyName,
                    PolicyDocument = policyDocument
                });
            policyArn = createPolicyResult.Policy.Arn;
        }
        catch (EntityAlreadyExistsException)
        {
            // The policy already exists, so we look it up to get the Arn.
            var policiesPaginator = _amazonIam.Paginators.ListPolicies(
                new ListPoliciesRequest()
                {
```

```
                    Scope = PolicyScopeType.Local
                });
            // Get the entire list using the paginator.
            await foreach (var policy in policiesPaginator.Policies)
            {
                if (policy.PolicyName.Equals(policyName))
                {
                    policyArn = policy.Arn;
                }
            }

            if (policyArn == null)
            {
                throw new InvalidOperationException("Policy not found");
            }
        }

        try
        {
            await _amazonIam.CreateRoleAsync(new CreateRoleRequest()
            {
                RoleName = roleName,
                AssumeRolePolicyDocument = assumeRoleDoc,
            });
            await _amazonIam.AttachRolePolicyAsync(new AttachRolePolicyRequest()
            {
                RoleName = roleName,
                PolicyArn = policyArn
            });
            if (awsManagedPolicies != null)
            {
                foreach (var awsPolicy in awsManagedPolicies)
                {
                    await _amazonIam.AttachRolePolicyAsync(new
AttachRolePolicyRequest()
                    {
                        PolicyArn = $"arn:aws:iam::aws:policy/{awsPolicy}",
                        RoleName = roleName
                    });
                }
            }
        }
        catch (EntityAlreadyExistsException)
        {
```

```
            Console.WriteLine("Role already exists.");
        }

        string profileArn = "";
        try
        {
            var profileCreateResponse = await
_amazonIam.CreateInstanceProfileAsync(
                new CreateInstanceProfileRequest()
                {
                    InstanceProfileName = profileName
                });
            // Allow time for the profile to be ready.
            profileArn = profileCreateResponse.InstanceProfile.Arn;
            Thread.Sleep(10000);
            await _amazonIam.AddRoleToInstanceProfileAsync(
                new AddRoleToInstanceProfileRequest()
                {
                    InstanceProfileName = profileName,
                    RoleName = roleName
                });

        }
        catch (EntityAlreadyExistsException)
        {
            Console.WriteLine("Policy already exists.");
            var profileGetResponse = await _amazonIam.GetInstanceProfileAsync(
                new GetInstanceProfileRequest()
                {
                    InstanceProfileName = profileName
                });
            profileArn = profileGetResponse.InstanceProfile.Arn;
        }
        return profileArn;
    }

    /// <summary>
    /// Create a new key pair and save the file.
    /// </summary>
    /// <param name="newKeyPairName">The name of the new key pair.</param>
    /// <returns>Async task.</returns>
    public async Task CreateKeyPair(string newKeyPairName)
    {
        try
```

```
        {
            var keyResponse = await _amazonEc2.CreateKeyPairAsync(
                new CreateKeyPairRequest() { KeyName = newKeyPairName });
            await File.WriteAllTextAsync($"{newKeyPairName}.pem",
                keyResponse.KeyPair.KeyMaterial);
            Console.WriteLine($"Created key pair {newKeyPairName}.");
        }
        catch (AlreadyExistsException)
        {
            Console.WriteLine("Key pair already exists.");
        }
    }

    /// <summary>
    /// Delete the key pair and file by name.
    /// </summary>
    /// <param name="deleteKeyPairName">The key pair to delete.</param>
    /// <returns>Async task.</returns>
    public async Task DeleteKeyPairByName(string deleteKeyPairName)
    {
        try
        {
            await _amazonEc2.DeleteKeyPairAsync(
                new DeleteKeyPairRequest() { KeyName = deleteKeyPairName });
            File.Delete($"{deleteKeyPairName}.pem");
        }
        catch (FileNotFoundException)
        {
            Console.WriteLine($"Key pair {deleteKeyPairName} not found.");
        }
    }

    /// <summary>
    /// Creates an Amazon EC2 launch template to use with Amazon EC2 Auto
Scaling.
    /// The launch template specifies a Bash script in its user data field that
runs after
    /// the instance is started. This script installs the Python packages and
starts a Python
    /// web server on the instance.
    /// </summary>
    /// <param name="startupScriptPath">The path to a Bash script file that is
run.</param>
```

```
   /// <param name="instancePolicyPath">The path to a permissions policy to
create and attach to the profile.</param>
   /// <returns>The template object.</returns>
   public async Task<Amazon.EC2.Model.LaunchTemplate> CreateTemplate(string
startupScriptPath, string instancePolicyPath)
   {
       try
       {
           await CreateKeyPair(_keyPairName);
           await CreateInstanceProfileWithName(_instancePolicyName,
_instanceRoleName,
               _instanceProfileName, instancePolicyPath);

           var startServerText = await File.ReadAllTextAsync(startupScriptPath);
           var plainTextBytes =
System.Text.Encoding.UTF8.GetBytes(startServerText);

           var amiLatest = await _amazonSsm.GetParameterAsync(
               new GetParameterRequest() { Name = _amiParam });
           var amiId = amiLatest.Parameter.Value;
           var launchTemplateResponse = await
_amazonEc2.CreateLaunchTemplateAsync(
               new CreateLaunchTemplateRequest()
               {
                   LaunchTemplateName = _launchTemplateName,
                   LaunchTemplateData = new RequestLaunchTemplateData()
                   {
                       InstanceType = _instanceType,
                       ImageId = amiId,
                       IamInstanceProfile =
                           new

LaunchTemplateIamInstanceProfileSpecificationRequest()
                           {
                               Name = _instanceProfileName
                           },
                       KeyName = _keyPairName,
                       UserData = System.Convert.ToBase64String(plainTextBytes)
                   }
               });
           return launchTemplateResponse.LaunchTemplate;
       }
       catch (AmazonEC2Exception ec2Exception)
       {
```

```
            if (ec2Exception.ErrorCode ==
"InvalidLaunchTemplateName.AlreadyExistsException")
            {
                _logger.LogError($"Could not create the template, the name
{_launchTemplateName} already exists. " +
                                 $"Please try again with a unique name.");
            }

            throw;
        }
        catch (Exception ex)
        {
            _logger.LogError($"An error occurred while creating the template.:
{ex.Message}");
            throw;
        }
    }

    /// <summary>
    /// Get a list of Availability Zones in the AWS Region of the Amazon EC2
Client.
    /// </summary>
    /// <returns>A list of availability zones.</returns>
    public async Task<List<string>> DescribeAvailabilityZones()
    {
        try
        {
            var zoneResponse = await _amazonEc2.DescribeAvailabilityZonesAsync(
                new DescribeAvailabilityZonesRequest());
            return zoneResponse.AvailabilityZones.Select(z =>
z.ZoneName).ToList();
        }
        catch (AmazonEC2Exception ec2Exception)
        {
            _logger.LogError($"An Amazon EC2 error occurred while listing
availability zones.: {ec2Exception.Message}");
            throw;
        }
        catch (Exception ex)
        {
            _logger.LogError($"An error occurred while listing availability
zones.: {ex.Message}");
            throw;
        }
```

```
    }

    /// <summary>
    /// Create an EC2 Auto Scaling group of a specified size and name.
    /// </summary>
    /// <param name="groupSize">The size for the group.</param>
    /// <param name="groupName">The name for the group.</param>
    /// <param name="availabilityZones">The availability zones for the group.</
param>
    /// <returns>Async task.</returns>
    public async Task CreateGroupOfSize(int groupSize, string groupName,
 List<string> availabilityZones)
    {
        try
        {
            await _amazonAutoScaling.CreateAutoScalingGroupAsync(
                new CreateAutoScalingGroupRequest()
                {
                    AutoScalingGroupName = groupName,
                    AvailabilityZones = availabilityZones,
                    LaunchTemplate =
                        new
 Amazon.AutoScaling.Model.LaunchTemplateSpecification()
                        {
                            LaunchTemplateName = _launchTemplateName,
                            Version = "$Default"
                        },
                    MaxSize = groupSize,
                    MinSize = groupSize
                });
            Console.WriteLine($"Created EC2 Auto Scaling group {groupName} with
 size {groupSize}.");
        }
        catch (EntityAlreadyExistsException)
        {
            Console.WriteLine($"EC2 Auto Scaling group {groupName} already
 exists.");
        }
    }

    /// <summary>
    /// Get the default VPC for the account.
    /// </summary>
    /// <returns>The default VPC object.</returns>
```

```
    public async Task<Vpc> GetDefaultVpc()
    {
        try
        {
            var vpcResponse = await _amazonEc2.DescribeVpcsAsync(
                new DescribeVpcsRequest()
                {
                    Filters = new List<Amazon.EC2.Model.Filter>()
                    {
                        new("is-default", new List<string>() { "true" })
                    }
                });
            return vpcResponse.Vpcs[0];
        }
        catch (AmazonEC2Exception ec2Exception)
        {
            if (ec2Exception.ErrorCode == "UnauthorizedOperation")
            {
                _logger.LogError(ec2Exception, $"You do not have the necessary
permissions to describe VPCs.");
            }

            throw;
        }
        catch (Exception ex)
        {
            _logger.LogError(ex, $"An error occurred while describing the vpcs.:
{ex.Message}");
            throw;
        }
    }

    /// <summary>
    /// Get all the subnets for a Vpc in a set of availability zones.
    /// </summary>
    /// <param name="vpcId">The Id of the Vpc.</param>
    /// <param name="availabilityZones">The list of availability zones.</param>
    /// <returns>The collection of subnet objects.</returns>
    public async Task<List<Subnet>> GetAllVpcSubnetsForZones(string vpcId,
List<string> availabilityZones)
    {
        try
        {
            var subnets = new List<Subnet>();
```

```
            var subnetPaginator = _amazonEc2.Paginators.DescribeSubnets(
                new DescribeSubnetsRequest()
                {
                    Filters = new List<Amazon.EC2.Model.Filter>()
                    {
                        new("vpc-id", new List<string>() { vpcId }),
                        new("availability-zone", availabilityZones),
                        new("default-for-az", new List<string>() { "true" })
                    }
                });

            // Get the entire list using the paginator.
            await foreach (var subnet in subnetPaginator.Subnets)
            {
                subnets.Add(subnet);
            }

            return subnets;
        }
        catch (AmazonEC2Exception ec2Exception)
        {
            if (ec2Exception.ErrorCode == "InvalidVpcID.NotFound")
            {
                _logger.LogError(ec2Exception, $"The specified VPC ID {vpcId}
does not exist.");
            }

            throw;
        }
        catch (Exception ex)
        {
            _logger.LogError(ex, $"An error occurred while describing the
subnets.: {ex.Message}");
            throw;
        }
    }

    /// <summary>
    /// Delete a launch template by name.
    /// </summary>
    /// <param name="templateName">The name of the template to delete.</param>
    /// <returns>Async task.</returns>
    public async Task DeleteTemplateByName(string templateName)
    {
```

```
        try
        {
            await _amazonEc2.DeleteLaunchTemplateAsync(
                new DeleteLaunchTemplateRequest()
                {
                    LaunchTemplateName = templateName
                });
        }
        catch (AmazonEC2Exception ec2Exception)
        {
            if (ec2Exception.ErrorCode ==
"InvalidLaunchTemplateName.NotFoundException")
            {
                _logger.LogError(
                    $"Could not delete the template, the name
{_launchTemplateName} was not found.");
            }

            throw;
        }
        catch (Exception ex)
        {
            _logger.LogError($"An error occurred while deleting the template.:
{ex.Message}");
            throw;
        }
    }

    /// <summary>
    /// Detaches a role from an instance profile, detaches policies from the
role,
    /// and deletes all the resources.
    /// </summary>
    /// <param name="profileName">The name of the profile to delete.</param>
    /// <param name="roleName">The name of the role to delete.</param>
    /// <returns>Async task.</returns>
    public async Task DeleteInstanceProfile(string profileName, string roleName)
    {
        try
        {
            await _amazonIam.RemoveRoleFromInstanceProfileAsync(
                new RemoveRoleFromInstanceProfileRequest()
                {
                    InstanceProfileName = profileName,
```

```
                    RoleName = roleName
                });
            await _amazonIam.DeleteInstanceProfileAsync(
                new DeleteInstanceProfileRequest() { InstanceProfileName =
profileName });
            var attachedPolicies = await
_amazonIam.ListAttachedRolePoliciesAsync(
                new ListAttachedRolePoliciesRequest() { RoleName = roleName });
            foreach (var policy in attachedPolicies.AttachedPolicies)
            {
                await _amazonIam.DetachRolePolicyAsync(
                    new DetachRolePolicyRequest()
                    {
                        RoleName = roleName,
                        PolicyArn = policy.PolicyArn
                    });
                // Delete the custom policies only.
                if (!policy.PolicyArn.StartsWith("arn:aws:iam::aws"))
                {
                    await _amazonIam.DeletePolicyAsync(
                        new Amazon.IdentityManagement.Model.DeletePolicyRequest()
                        {
                            PolicyArn = policy.PolicyArn
                        });
                }
            }

            await _amazonIam.DeleteRoleAsync(
                new DeleteRoleRequest() { RoleName = roleName });
        }
        catch (NoSuchEntityException)
        {
            Console.WriteLine($"Instance profile {profileName} does not exist.");
        }
    }

    /// <summary>
    /// Gets data about the instances in an EC2 Auto Scaling group by its group
name.
    /// </summary>
    /// <param name="group">The name of the auto scaling group.</param>
    /// <returns>A collection of instance Ids.</returns>
    public async Task<IEnumerable<string>> GetInstancesByGroupName(string group)
    {
```

```
        var instanceResponse = await
_amazonAutoScaling.DescribeAutoScalingGroupsAsync(
            new DescribeAutoScalingGroupsRequest()
            {
                AutoScalingGroupNames = new List<string>() { group }
            });
        var instanceIds = instanceResponse.AutoScalingGroups.SelectMany(
            g => g.Instances.Select(i => i.InstanceId));
        return instanceIds;
    }

    /// <summary>
    /// Get the instance profile association data for an instance.
    /// </summary>
    /// <param name="instanceId">The Id of the instance.</param>
    /// <returns>Instance profile associations data.</returns>
    public async Task<IamInstanceProfileAssociation> GetInstanceProfile(string
instanceId)
    {
        try
        {
            var response = await
_amazonEc2.DescribeIamInstanceProfileAssociationsAsync(
                new DescribeIamInstanceProfileAssociationsRequest()
                {
                    Filters = new List<Amazon.EC2.Model.Filter>()
                    {
                        new("instance-id", new List<string>() { instanceId })
                    },
                });
            return response.IamInstanceProfileAssociations[0];
        }
        catch (AmazonEC2Exception ec2Exception)
        {
            if (ec2Exception.ErrorCode == "InvalidInstanceID.NotFound")
            {
                _logger.LogError(ec2Exception, $"Instance {instanceId} not
found");
            }

            throw;
        }
        catch (Exception ex)
        {
```

```
            _logger.LogError(ex, $"An error occurred while creating the
template.: {ex.Message}");
            throw;
        }
    }

    /// <summary>
    /// Replace the profile associated with a running instance. After the profile
is replaced, the instance
    /// is rebooted to ensure that it uses the new profile. When the instance is
ready, Systems Manager is
    /// used to restart the Python web server.
    /// </summary>
    /// <param name="instanceId">The Id of the instance to update.</param>
    /// <param name="credsProfileName">The name of the new profile to associate
with the specified instance.</param>
    /// <param name="associationId">The Id of the existing profile association
for the instance.</param>
    /// <returns>Async task.</returns>
    public async Task ReplaceInstanceProfile(string instanceId, string
credsProfileName, string associationId)
    {
        try
        {
            await _amazonEc2.ReplaceIamInstanceProfileAssociationAsync(
                new ReplaceIamInstanceProfileAssociationRequest()
                {
                    AssociationId = associationId,
                    IamInstanceProfile = new IamInstanceProfileSpecification()
                    {
                        Name = credsProfileName
                    }
                });
            // Allow time before resetting.
            Thread.Sleep(25000);

            await _amazonEc2.RebootInstancesAsync(
                new RebootInstancesRequest(new List<string>() { instanceId }));
            Thread.Sleep(25000);
            var instanceReady = false;
            var retries = 5;
            while (retries-- > 0 && !instanceReady)
            {
                var instancesPaginator =
```

```
                    _amazonSsm.Paginators.DescribeInstanceInformation(
                        new DescribeInstanceInformationRequest());
                // Get the entire list using the paginator.
                await foreach (var instance in
instancesPaginator.InstanceInformationList)
                {
                    instanceReady = instance.InstanceId == instanceId;
                    if (instanceReady)
                    {
                        break;
                    }
                }
            }
            Console.WriteLine("Waiting for instance to be running.");
            await WaitForInstanceState(instanceId, InstanceStateName.Running);
            Console.WriteLine("Instance ready.");
            Console.WriteLine($"Sending restart command to instance
{instanceId}");
            await _amazonSsm.SendCommandAsync(
                new SendCommandRequest()
                {
                    InstanceIds = new List<string>() { instanceId },
                    DocumentName = "AWS-RunShellScript",
                    Parameters = new Dictionary<string, List<string>>()
                    {
                        {
                            "commands",
                            new List<string>() { "cd / && sudo python3 server.py
80" }
                        }
                    }
                });
            Console.WriteLine($"Restarted the web server on instance
{instanceId}");
        }
        catch (AmazonEC2Exception ec2Exception)
        {
            if (ec2Exception.ErrorCode == "InvalidInstanceID.NotFound")
            {
                _logger.LogError(ec2Exception, $"Instance {instanceId} not
found");
            }

            throw;
```

```
        }
        catch (Exception ex)
        {
            _logger.LogError(ex, $"An error occurred while replacing the
template.: {ex.Message}");
            throw;
        }
    }

    /// <summary>
    /// Try to terminate an instance by its Id.
    /// </summary>
    /// <param name="instanceId">The Id of the instance to terminate.</param>
    /// <returns>Async task.</returns>
    public async Task TryTerminateInstanceById(string instanceId)
    {
        var stopping = false;
        Console.WriteLine($"Stopping {instanceId}...");
        while (!stopping)
        {
            try
            {
                await
_amazonAutoScaling.TerminateInstanceInAutoScalingGroupAsync(
                    new TerminateInstanceInAutoScalingGroupRequest()
                    {
                        InstanceId = instanceId,
                        ShouldDecrementDesiredCapacity = false
                    });
                stopping = true;
            }
            catch (ScalingActivityInProgressException)
            {
                Console.WriteLine($"Scaling activity in progress for
{instanceId}. Waiting...");
                Thread.Sleep(10000);
            }
        }
    }

    /// <summary>
    /// Tries to delete the EC2 Auto Scaling group. If the group is in use or in
progress,
    /// waits and retries until the group is successfully deleted.
```

```
    /// </summary>
    /// <param name="groupName">The name of the group to try to delete.</param>
    /// <returns>Async task.</returns>
    public async Task TryDeleteGroupByName(string groupName)
    {
        var stopped = false;
        while (!stopped)
        {
            try
            {
                await _amazonAutoScaling.DeleteAutoScalingGroupAsync(
                    new DeleteAutoScalingGroupRequest()
                    {
                        AutoScalingGroupName = groupName
                    });
                stopped = true;
            }
            catch (Exception e)
                when ((e is ScalingActivityInProgressException)
                      || (e is Amazon.AutoScaling.Model.ResourceInUseException))
            {
                Console.WriteLine($"Some instances are still running.
Waiting...");
                Thread.Sleep(10000);
            }
        }
    }

    /// <summary>
    /// Terminate instances and delete the Auto Scaling group by name.
    /// </summary>
    /// <param name="groupName">The name of the group to delete.</param>
    /// <returns>Async task.</returns>
    public async Task TerminateAndDeleteAutoScalingGroupWithName(string
groupName)
    {
        var describeGroupsResponse = await
_amazonAutoScaling.DescribeAutoScalingGroupsAsync(
            new DescribeAutoScalingGroupsRequest()
            {
                AutoScalingGroupNames = new List<string>() { groupName }
            });
        if (describeGroupsResponse.AutoScalingGroups.Any())
        {
```

```
            // Update the size to 0.
            await _amazonAutoScaling.UpdateAutoScalingGroupAsync(
                new UpdateAutoScalingGroupRequest()
                {
                    AutoScalingGroupName = groupName,
                    MinSize = 0
                });
            var group = describeGroupsResponse.AutoScalingGroups[0];
            foreach (var instance in group.Instances)
            {
                await TryTerminateInstanceById(instance.InstanceId);
            }

            await TryDeleteGroupByName(groupName);
        }
        else
        {
            Console.WriteLine($"No groups found with name {groupName}.");
        }
    }


    /// <summary>
    /// Get the default security group for a specified Vpc.
    /// </summary>
    /// <param name="vpc">The Vpc to search.</param>
    /// <returns>The default security group.</returns>
    public async Task<SecurityGroup> GetDefaultSecurityGroupForVpc(Vpc vpc)
    {
        var groupResponse = await _amazonEc2.DescribeSecurityGroupsAsync(
            new DescribeSecurityGroupsRequest()
            {
                Filters = new List<Amazon.EC2.Model.Filter>()
                {
                    new ("group-name", new List<string>() { "default" }),
                    new ("vpc-id", new List<string>() { vpc.VpcId })
                }
            });
        return groupResponse.SecurityGroups[0];
    }

    /// <summary>
    /// Verify the default security group of a Vpc allows ingress from the
calling computer.
```

```
   /// This can be done by allowing ingress from this computer's IP address.
   /// In some situations, such as connecting from a corporate network, you must
instead specify
   /// a prefix list Id. You can also temporarily open the port to any IP
address while running this example.
   /// If you do, be sure to remove public access when you're done.
   /// </summary>
   /// <param name="vpc">The group to check.</param>
   /// <param name="port">The port to verify.</param>
   /// <param name="ipAddress">This computer's IP address.</param>
   /// <returns>True if the ip address is allowed on the group.</returns>
   public bool VerifyInboundPortForGroup(SecurityGroup group, int port, string
ipAddress)
   {
       var portIsOpen = false;
       foreach (var ipPermission in group.IpPermissions)
       {
           if (ipPermission.FromPort == port)
           {
               foreach (var ipRange in ipPermission.Ipv4Ranges)
               {
                   var cidr = ipRange.CidrIp;
                   if (cidr.StartsWith(ipAddress) || cidr == "0.0.0.0/0")
                   {
                       portIsOpen = true;
                   }
               }

               if (ipPermission.PrefixListIds.Any())
               {
                   portIsOpen = true;
               }

               if (!portIsOpen)
               {
                   Console.WriteLine("The inbound rule does not appear to be
open to either this computer's IP\n" +
                                     "address, to all IP addresses (0.0.0.0/0),
or to a prefix list ID.");
               }
               else
               {
                   break;
               }
```

```
            }
        }

        return portIsOpen;
    }

    /// <summary>
    /// Add an ingress rule to the specified security group that allows access on
 the
    /// specified port from the specified IP address.
    /// </summary>
    /// <param name="groupId">The Id of the security group to modify.</param>
    /// <param name="port">The port to open.</param>
    /// <param name="ipAddress">The IP address to allow access.</param>
    /// <returns>Async task.</returns>
    public async Task OpenInboundPort(string groupId, int port, string ipAddress)
    {
        await _amazonEc2.AuthorizeSecurityGroupIngressAsync(
            new AuthorizeSecurityGroupIngressRequest()
            {
                GroupId = groupId,
                IpPermissions = new List<IpPermission>()
                {
                    new IpPermission()
                    {
                        FromPort = port,
                        ToPort = port,
                        IpProtocol = "tcp",
                        Ipv4Ranges = new List<IpRange>()
                        {
                            new IpRange() { CidrIp = $"{ipAddress}/32" }
                        }
                    }
                }
            });
    }

    /// <summary>
    /// Attaches an Elastic Load Balancing (ELB) target group to this EC2 Auto
 Scaling group.
    /// The
    /// </summary>
    /// <param name="autoScalingGroupName">The name of the Auto Scaling group.</
param>
```

```
    /// <param name="targetGroupArn">The Arn for the target group.</param>
    /// <returns>Async task.</returns>
    public async Task AttachLoadBalancerToGroup(string autoScalingGroupName,
 string targetGroupArn)
    {
        await _amazonAutoScaling.AttachLoadBalancerTargetGroupsAsync(
            new AttachLoadBalancerTargetGroupsRequest()
            {
                AutoScalingGroupName = autoScalingGroupName,
                TargetGroupARNs = new List<string>() { targetGroupArn }
            });
    }

    /// <summary>
    /// Wait until an EC2 instance is in a specified state.
    /// </summary>
    /// <param name="instanceId">The instance Id.</param>
    /// <param name="stateName">The state to wait for.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> WaitForInstanceState(string instanceId,
 InstanceStateName stateName)
    {
        var request = new DescribeInstancesRequest
        {
            InstanceIds = new List<string> { instanceId }
        };

        // Wait until the instance is in the specified state.
        var hasState = false;
        do
        {
            // Wait 5 seconds.
            Thread.Sleep(5000);

            // Check for the desired state.
            var response = await _amazonEc2.DescribeInstancesAsync(request);
            var instance = response.Reservations[0].Instances[0];
            hasState = instance.State.Name == stateName;
            Console.Write(". ");
        } while (!hasState);

        return hasState;
    }
}
```

Crie uma classe que envolva ações do Elastic Load Balancing.

```
/// <summary>
/// Encapsulates Elastic Load Balancer actions.
/// </summary>
public class ElasticLoadBalancerWrapper
{
    private readonly IAmazonElasticLoadBalancingV2 _amazonElasticLoadBalancingV2;
    private string? _endpoint = null;
    private readonly string _targetGroupName = "";
    private readonly string _loadBalancerName = "";
    HttpClient _httpClient = new();

    public string TargetGroupName => _targetGroupName;
    public string LoadBalancerName => _loadBalancerName;

    /// <summary>
    /// Constructor for the Elastic Load Balancer wrapper.
    /// </summary>
    /// <param name="amazonElasticLoadBalancingV2">The injected load balancing v2
 client.</param>
    /// <param name="configuration">The injected configuration.</param>
    public ElasticLoadBalancerWrapper(
        IAmazonElasticLoadBalancingV2 amazonElasticLoadBalancingV2,
        IConfiguration configuration)
    {
        _amazonElasticLoadBalancingV2 = amazonElasticLoadBalancingV2;
        var prefix = configuration["resourcePrefix"];
        _targetGroupName = prefix + "-tg";
        _loadBalancerName = prefix + "-lb";
    }

    /// <summary>
    /// Get the HTTP Endpoint of a load balancer by its name.
    /// </summary>
    /// <param name="loadBalancerName">The name of the load balancer.</param>
    /// <returns>The HTTP endpoint.</returns>
    public async Task<string> GetEndpointForLoadBalancerByName(string
 loadBalancerName)
    {
```

```
        if (_endpoint == null)
        {
            var endpointResponse =
                await _amazonElasticLoadBalancingV2.DescribeLoadBalancersAsync(
                    new DescribeLoadBalancersRequest()
                    {
                        Names = new List<string>() { loadBalancerName }
                    });
            _endpoint = endpointResponse.LoadBalancers[0].DNSName;
        }

        return _endpoint;
    }

    /// <summary>
    /// Return the GET response for an endpoint as text.
    /// </summary>
    /// <param name="endpoint">The endpoint for the request.</param>
    /// <returns>The request response.</returns>
    public async Task<string> GetEndPointResponse(string endpoint)
    {
        var endpointResponse = await _httpClient.GetAsync($"http://{endpoint}");
        var textResponse = await endpointResponse.Content.ReadAsStringAsync();
        return textResponse!;
    }

    /// <summary>
    /// Get the target health for a group by name.
    /// </summary>
    /// <param name="groupName">The name of the group.</param>
    /// <returns>The collection of health descriptions.</returns>
    public async Task<List<TargetHealthDescription>>
CheckTargetHealthForGroup(string groupName)
    {
        List<TargetHealthDescription> result = null!;
        try
        {
            var groupResponse =
                await _amazonElasticLoadBalancingV2.DescribeTargetGroupsAsync(
                    new DescribeTargetGroupsRequest()
                    {
                        Names = new List<string>() { groupName }
                    });
            var healthResponse =
```

```
                await _amazonElasticLoadBalancingV2.DescribeTargetHealthAsync(
                    new DescribeTargetHealthRequest()
                    {
                        TargetGroupArn =
 groupResponse.TargetGroups[0].TargetGroupArn
                    });
            ;
            result = healthResponse.TargetHealthDescriptions;
        }
        catch (TargetGroupNotFoundException)
        {
            Console.WriteLine($"Target group {groupName} not found.");
        }
        return result;
    }

    /// <summary>
    /// Create an Elastic Load Balancing target group. The target group specifies
 how the load balancer forwards
    /// requests to instances in the group and how instance health is checked.
    ///
    /// To speed up this demo, the health check is configured with shortened
 times and lower thresholds. In production,
    /// you might want to decrease the sensitivity of your health checks to avoid
 unwanted failures.
    /// </summary>
    /// <param name="groupName">The name for the group.</param>
    /// <param name="protocol">The protocol, such as HTTP.</param>
    /// <param name="port">The port to use to forward requests, such as 80.</
param>
    /// <param name="vpcId">The Id of the Vpc in which the load balancer
 exists.</param>
    /// <returns>The new TargetGroup object.</returns>
    public async Task<TargetGroup> CreateTargetGroupOnVpc(string groupName,
 ProtocolEnum protocol, int port, string vpcId)
    {
        var createResponse = await
 _amazonElasticLoadBalancingV2.CreateTargetGroupAsync(
            new CreateTargetGroupRequest()
            {
                Name = groupName,
                Protocol = protocol,
                Port = port,
                HealthCheckPath = "/healthcheck",
```

```
                HealthCheckIntervalSeconds = 10,
                HealthCheckTimeoutSeconds = 5,
                HealthyThresholdCount = 2,
                UnhealthyThresholdCount = 2,
                VpcId = vpcId
            });
        var targetGroup = createResponse.TargetGroups[0];
        return targetGroup;
    }

    /// <summary>
    /// Create an Elastic Load Balancing load balancer that uses the specified
subnets
    /// and forwards requests to the specified target group.
    /// </summary>
    /// <param name="name">The name for the new load balancer.</param>
    /// <param name="subnetIds">Subnets for the load balancer.</param>
    /// <param name="targetGroup">Target group for forwarded requests.</param>
    /// <returns>The new LoadBalancer object.</returns>
    public async Task<LoadBalancer> CreateLoadBalancerAndListener(string name,
List<string> subnetIds, TargetGroup targetGroup)
    {
        var createLbResponse = await
_amazonElasticLoadBalancingV2.CreateLoadBalancerAsync(
            new CreateLoadBalancerRequest()
            {
                Name = name,
                Subnets = subnetIds
            });
        var loadBalancerArn = createLbResponse.LoadBalancers[0].LoadBalancerArn;

        // Wait for load balancer to be available.
        var loadBalancerReady = false;
        while (!loadBalancerReady)
        {
            try
            {
                var describeResponse =
                    await
_amazonElasticLoadBalancingV2.DescribeLoadBalancersAsync(
                        new DescribeLoadBalancersRequest()
                        {
                            Names = new List<string>() { name }
                        });
```

```
                var loadBalancerState =
describeResponse.LoadBalancers[0].State.Code;

                loadBalancerReady = loadBalancerState ==
LoadBalancerStateEnum.Active;
            }
            catch (LoadBalancerNotFoundException)
            {
                loadBalancerReady = false;
            }
            Thread.Sleep(10000);
        }
        // Create the listener.
        await _amazonElasticLoadBalancingV2.CreateListenerAsync(
            new CreateListenerRequest()
            {
                LoadBalancerArn = loadBalancerArn,
                Protocol = targetGroup.Protocol,
                Port = targetGroup.Port,
                DefaultActions = new List<Action>()
                {
                    new Action()
                    {
                        Type = ActionTypeEnum.Forward,
                        TargetGroupArn = targetGroup.TargetGroupArn
                    }
                }
            });
        return createLbResponse.LoadBalancers[0];
    }

    /// <summary>
    /// Verify this computer can successfully send a GET request to the
    /// load balancer endpoint.
    /// </summary>
    /// <param name="endpoint">The endpoint to check.</param>
    /// <returns>True if successful.</returns>
    public async Task<bool> VerifyLoadBalancerEndpoint(string endpoint)
    {
        var success = false;
        var retries = 3;
        while (!success && retries > 0)
        {
```

```
            try
            {
                var endpointResponse = await _httpClient.GetAsync($"http://
{endpoint}");
                Console.WriteLine($"Response: {endpointResponse.StatusCode}.");

                if (endpointResponse.IsSuccessStatusCode)
                {
                    success = true;
                }
                else
                {
                    retries = 0;
                }
            }
            catch (HttpRequestException)
            {
                Console.WriteLine("Connection error, retrying...");
                retries--;
                Thread.Sleep(10000);
            }
        }

        return success;
    }

    /// <summary>
    /// Delete a load balancer by its specified name.
    /// </summary>
    /// <param name="name">The name of the load balancer to delete.</param>
    /// <returns>Async task.</returns>
    public async Task DeleteLoadBalancerByName(string name)
    {
        try
        {
            var describeLoadBalancerResponse =
                await _amazonElasticLoadBalancingV2.DescribeLoadBalancersAsync(
                    new DescribeLoadBalancersRequest()
                    {
                        Names = new List<string>() { name }
                    });
            var lbArn =
 describeLoadBalancerResponse.LoadBalancers[0].LoadBalancerArn;
            await _amazonElasticLoadBalancingV2.DeleteLoadBalancerAsync(
```

```
                new DeleteLoadBalancerRequest()
                {
                    LoadBalancerArn = lbArn
                }
            );
        }
        catch (LoadBalancerNotFoundException)
        {
            Console.WriteLine($"Load balancer {name} not found.");
        }
    }

    /// <summary>
    /// Delete a TargetGroup by its specified name.
    /// </summary>
    /// <param name="groupName">Name of the group to delete.</param>
    /// <returns>Async task.</returns>
    public async Task DeleteTargetGroupByName(string groupName)
    {
        var done = false;
        while (!done)
        {
            try
            {
                var groupResponse =
                    await
_amazonElasticLoadBalancingV2.DescribeTargetGroupsAsync(
                        new DescribeTargetGroupsRequest()
                        {
                            Names = new List<string>() { groupName }
                        });

                var targetArn = groupResponse.TargetGroups[0].TargetGroupArn;
                await _amazonElasticLoadBalancingV2.DeleteTargetGroupAsync(
                    new DeleteTargetGroupRequest() { TargetGroupArn =
targetArn });
                Console.WriteLine($"Deleted load balancing target group
{groupName}.");
                done = true;
            }
            catch (TargetGroupNotFoundException)
            {
                Console.WriteLine(
                    $"Target group {groupName} not found, could not delete.");
```

```
                done = true;
            }
            catch (ResourceInUseException)
            {
                Console.WriteLine("Target group not yet released, waiting...");
                Thread.Sleep(10000);
            }
        }
    }
}
```

Crie uma classe que use o DynamoDB para simular um serviço de recomendação.

```
/// <summary>
/// Encapsulates a DynamoDB table to use as a service that recommends books,
 movies, and songs.
/// </summary>
public class Recommendations
{
    private readonly IAmazonDynamoDB _amazonDynamoDb;
    private readonly DynamoDBContext _context;
    private readonly string _tableName;

    public string TableName => _tableName;

    /// <summary>
    /// Constructor for the Recommendations service.
    /// </summary>
    /// <param name="amazonDynamoDb">The injected DynamoDb client.</param>
    /// <param name="configuration">The injected configuration.</param>
    public Recommendations(IAmazonDynamoDB amazonDynamoDb, IConfiguration
 configuration)
    {
        _amazonDynamoDb = amazonDynamoDb;
        _context = new DynamoDBContext(_amazonDynamoDb);
        _tableName = configuration["databaseName"]!;
    }

    /// <summary>
    /// Create the DynamoDb table with a specified name.
    /// </summary>
    /// <param name="tableName">The name for the table.</param>
```

```
    /// <returns>True when ready.</returns>
    public async Task<bool> CreateDatabaseWithName(string tableName)
    {
        try
        {
            Console.Write($"Creating table {tableName}...");
            var createRequest = new CreateTableRequest()
            {
                TableName = tableName,
                AttributeDefinitions = new List<AttributeDefinition>()
                    {
                        new AttributeDefinition()
                        {
                            AttributeName = "MediaType",
                            AttributeType = ScalarAttributeType.S
                        },
                        new AttributeDefinition()
                        {
                            AttributeName = "ItemId",
                            AttributeType = ScalarAttributeType.N
                        }
                    },
                KeySchema = new List<KeySchemaElement>()
                    {
                        new KeySchemaElement()
                        {
                            AttributeName = "MediaType",
                            KeyType = KeyType.HASH
                        },
                        new KeySchemaElement()
                        {
                            AttributeName = "ItemId",
                            KeyType = KeyType.RANGE
                        }
                    },
                ProvisionedThroughput = new ProvisionedThroughput()
                {
                    ReadCapacityUnits = 5,
                    WriteCapacityUnits = 5
                }
            };
            await _amazonDynamoDb.CreateTableAsync(createRequest);

            // Wait until the table is ACTIVE and then report success.
```

```
            Console.Write("\nWaiting for table to become active...");

            var request = new DescribeTableRequest
            {
                TableName = tableName
            };

            TableStatus status;
            do
            {
                Thread.Sleep(2000);

                var describeTableResponse = await
 _amazonDynamoDb.DescribeTableAsync(request);
                status = describeTableResponse.Table.TableStatus;

                Console.Write(".");
            }
            while (status != "ACTIVE");

            return status == TableStatus.ACTIVE;
        }
        catch (ResourceInUseException)
        {
            Console.WriteLine($"Table {tableName} already exists.");
            return false;
        }
    }

    /// <summary>
    /// Populate the database table with data from a specified path.
    /// </summary>
    /// <param name="databaseTableName">The name of the table.</param>
    /// <param name="recommendationsPath">The path of the recommendations data.</
param>
    /// <returns>Async task.</returns>
    public async Task PopulateDatabase(string databaseTableName, string
 recommendationsPath)
    {
        var recommendationsText = await
 File.ReadAllTextAsync(recommendationsPath);
        var records =

 JsonSerializer.Deserialize<RecommendationModel[]>(recommendationsText);
```

```
        var batchWrite = _context.CreateBatchWrite<RecommendationModel>();

        foreach (var record in records!)
        {
            batchWrite.AddPutItem(record);
        }

        await batchWrite.ExecuteAsync();
    }

    /// <summary>
    /// Delete the recommendation table by name.
    /// </summary>
    /// <param name="tableName">The name of the recommendation table.</param>
    /// <returns>Async task.</returns>
    public async Task DestroyDatabaseByName(string tableName)
    {
        try
        {
            await _amazonDynamoDb.DeleteTableAsync(
                new DeleteTableRequest() { TableName = tableName });
            Console.WriteLine($"Table {tableName} was deleted.");
        }
        catch (ResourceNotFoundException)
        {
            Console.WriteLine($"Table {tableName} not found");
        }
    }
}
```

Crie uma classe que envolva as ações do Systems Manager.

```
/// <summary>
/// Encapsulates Systems Manager parameter operations. This example uses these
 parameters
/// to drive the demonstration of resilient architecture, such as failure of a
 dependency or
/// how the service responds to a health check.
/// </summary>
public class SmParameterWrapper
{
```

```
    private readonly IAmazonSimpleSystemsManagement
 _amazonSimpleSystemsManagement;

    private readonly string _tableParameter = "doc-example-resilient-
architecture-table";
    private readonly string _failureResponseParameter = "doc-example-resilient-
architecture-failure-response";
    private readonly string _healthCheckParameter = "doc-example-resilient-
architecture-health-check";
    private readonly string _tableName = "";

    public string TableParameter => _tableParameter;
    public string TableName => _tableName;
    public string HealthCheckParameter => _healthCheckParameter;
    public string FailureResponseParameter => _failureResponseParameter;

    /// <summary>
    /// Constructor for the SmParameterWrapper.
    /// </summary>
    /// <param name="amazonSimpleSystemsManagement">The injected Simple Systems
 Management client.</param>
    /// <param name="configuration">The injected configuration.</param>
    public SmParameterWrapper(IAmazonSimpleSystemsManagement
 amazonSimpleSystemsManagement, IConfiguration configuration)
    {
        _amazonSimpleSystemsManagement = amazonSimpleSystemsManagement;
        _tableName = configuration["databaseName"]!;
    }

    /// <summary>
    /// Reset the Systems Manager parameters to starting values for the demo.
    /// </summary>
    /// <returns>Async task.</returns>
    public async Task Reset()
    {
        await this.PutParameterByName(_tableParameter, _tableName);
        await this.PutParameterByName(_failureResponseParameter, "none");
        await this.PutParameterByName(_healthCheckParameter, "shallow");
    }

    /// <summary>
    /// Set the value of a named Systems Manager parameter.
    /// </summary>
    /// <param name="name">The name of the parameter.</param>
```

```
    /// <param name="value">The value to set.</param>
    /// <returns>Async task.</returns>
    public async Task PutParameterByName(string name, string value)
    {
        await _amazonSimpleSystemsManagement.PutParameterAsync(
            new PutParameterRequest() { Name = name, Value = value, Overwrite =
 true });
    }
}
```

- Para obter detalhes da API, consulte os tópicos a seguir na Referência da API AWS SDK para .NET .
  - AttachLoadBalancerTargetGroups
  - CreateAutoScalingGroup
  - CreateInstanceProfile
  - CreateLaunchTemplate
  - CreateListener
  - CreateLoadBalancer
  - CreateTargetGroup
  - DeleteAutoScalingGroup
  - DeleteInstanceProfile
  - DeleteLaunchTemplate
  - DeleteLoadBalancer
  - DeleteTargetGroup
  - DescribeAutoScalingGroups
  - DescribeAvailabilityZones
  - DescribeIamInstanceProfileAssociations
  - DescribeInstances
  - DescribeLoadBalancers
  - DescribeSubnets
  - DescribeTargetGroups
  - DescribeTargetHealth
  - DescribeVpcs

- RebootInstances
- ReplaceIamInstanceProfileAssociation
- TerminateInstanceInAutoScalingGroup
- UpdateAutoScalingGroup

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

Execute o cenário interativo em um prompt de comando.

```
public class Main {

    public static final String fileName = "C:\\AWS\\resworkflow\
\recommendations.json"; // Modify file location.
    public static final String tableName = "doc-example-recommendation-service";
    public static final String startScript = "C:\\AWS\\resworkflow\
\server_startup_script.sh"; // Modify file location.
    public static final String policyFile = "C:\\AWS\\resworkflow\
\instance_policy.json"; // Modify file location.
    public static final String ssmJSON = "C:\\AWS\\resworkflow\
\ssm_only_policy.json"; // Modify file location.
    public static final String failureResponse = "doc-example-resilient-
architecture-failure-response";
    public static final String healthCheck = "doc-example-resilient-architecture-
health-check";
    public static final String templateName = "doc-example-resilience-template";
    public static final String roleName = "doc-example-resilience-role";
    public static final String policyName = "doc-example-resilience-pol";
    public static final String profileName = "doc-example-resilience-prof";

    public static final String badCredsProfileName = "doc-example-resilience-
prof-bc";
```

```
    public static final String targetGroupName = "doc-example-resilience-tg";
    public static final String autoScalingGroupName = "doc-example-resilience-
group";
    public static final String lbName = "doc-example-resilience-lb";
    public static final String protocol = "HTTP";
    public static final int port = 80;

    public static final String DASHES = new String(new char[80]).replace("\0",
 "-");

    public static void main(String[] args) throws IOException,
 InterruptedException {
        Scanner in = new Scanner(System.in);
        Database database = new Database();
        AutoScaler autoScaler = new AutoScaler();
        LoadBalancer loadBalancer = new LoadBalancer();

        System.out.println(DASHES);
        System.out.println("Welcome to the demonstration of How to Build and
 Manage a Resilient Service!");
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("A - SETUP THE RESOURCES");
        System.out.println("Press Enter when you're ready to start deploying
 resources.");
        in.nextLine();
        deploy(loadBalancer);
        System.out.println(DASHES);
        System.out.println(DASHES);
        System.out.println("B - DEMO THE RESILIENCE FUNCTIONALITY");
        System.out.println("Press Enter when you're ready.");
        in.nextLine();
        demo(loadBalancer);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("C - DELETE THE RESOURCES");
        System.out.println("""
                This concludes the demo of how to build and manage a resilient
 service.
                To keep things tidy and to avoid unwanted charges on your
 account, we can clean up all AWS resources
                that were created for this demo.
```

```
                """);

        System.out.println("\n Do you want to delete the resources (y/n)? ");
        String userInput = in.nextLine().trim().toLowerCase(); // Capture user
input

        if (userInput.equals("y")) {
            // Delete resources here
            deleteResources(loadBalancer, autoScaler, database);
            System.out.println("Resources deleted.");
        } else {
            System.out.println("""
                    Okay, we'll leave the resources intact.
                    Don't forget to delete them when you're done with them or you
might incur unexpected charges.
                    """);
        }
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("The example has completed. ");
        System.out.println("\n Thanks for watching!");
        System.out.println(DASHES);
    }

    // Deletes the AWS resources used in this example.
    private static void deleteResources(LoadBalancer loadBalancer, AutoScaler
autoScaler, Database database)
            throws IOException, InterruptedException {
        loadBalancer.deleteLoadBalancer(lbName);
        System.out.println("*** Wait 30 secs for resource to be deleted");
        TimeUnit.SECONDS.sleep(30);
        loadBalancer.deleteTargetGroup(targetGroupName);
        autoScaler.deleteAutoScaleGroup(autoScalingGroupName);
        autoScaler.deleteRolesPolicies(policyName, roleName, profileName);
        autoScaler.deleteTemplate(templateName);
        database.deleteTable(tableName);
    }

    private static void deploy(LoadBalancer loadBalancer) throws
InterruptedException, IOException {
        Scanner in = new Scanner(System.in);
        System.out.println(
                """
```

```
                        For this demo, we'll use the AWS SDK for Java (v2) to
create several AWS resources
                        to set up a load-balanced web service endpoint and
explore some ways to make it resilient
                        against various kinds of failures.

                        Some of the resources create by this demo are:
                        \t* A DynamoDB table that the web service depends on to
provide book, movie, and song recommendations.
                        \t* An EC2 launch template that defines EC2 instances
that each contain a Python web server.
                        \t* An EC2 Auto Scaling group that manages EC2 instances
across several Availability Zones.
                        \t* An Elastic Load Balancing (ELB) load balancer that
targets the Auto Scaling group to distribute requests.
                        """);

        System.out.println("Press Enter when you're ready.");
        in.nextLine();
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("Creating and populating a DynamoDB table named " +
tableName);
        Database database = new Database();
        database.createTable(tableName, fileName);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("""
                Creating an EC2 launch template that runs '{startup_script}' when
an instance starts.
                This script starts a Python web server defined in the `server.py`
script. The web server
                listens to HTTP requests on port 80 and responds to requests to
'/' and to '/healthcheck'.
                For demo purposes, this server is run as the root user. In
production, the best practice is to
                run a web server, such as Apache, with least-privileged
credentials.

                The template also defines an IAM policy that each instance uses
to assume a role that grants
```

```
                permissions to access the DynamoDB recommendation table and
Systems Manager parameters
                that control the flow of the demo.
                """);

        LaunchTemplateCreator templateCreator = new LaunchTemplateCreator();
        templateCreator.createTemplate(policyFile, policyName, profileName,
startScript, templateName, roleName);
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println(
                "Creating an EC2 Auto Scaling group that maintains three EC2
instances, each in a different Availability Zone.");
        System.out.println("*** Wait 30 secs for the VPC to be created");
        TimeUnit.SECONDS.sleep(30);
        AutoScaler autoScaler = new AutoScaler();
        String[] zones = autoScaler.createGroup(3, templateName,
autoScalingGroupName);

        System.out.println("""
                At this point, you have EC2 instances created. Once each instance
starts, it listens for
                HTTP requests. You can see these instances in the console or
continue with the demo.
                Press Enter when you're ready to continue.
                """);

        in.nextLine();
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("Creating variables that control the flow of the
demo.");
        ParameterHelper paramHelper = new ParameterHelper();
        paramHelper.reset();
        System.out.println(DASHES);

        System.out.println(DASHES);
        System.out.println("""
                Creating an Elastic Load Balancing target group and load
balancer. The target group
                defines how the load balancer connects to instances. The load
balancer provides a
```

```
                single endpoint where clients connect and dispatches requests to
instances in the group.
                """);

        String vpcId = autoScaler.getDefaultVPC();
        List<Subnet> subnets = autoScaler.getSubnets(vpcId, zones);
        System.out.println("You have retrieved a list with " + subnets.size() + "
subnets");
        String targetGroupArn = loadBalancer.createTargetGroup(protocol, port,
vpcId, targetGroupName);
        String elbDnsName = loadBalancer.createLoadBalancer(subnets,
targetGroupArn, lbName, port, protocol);
        autoScaler.attachLoadBalancerTargetGroup(autoScalingGroupName,
targetGroupArn);
        System.out.println("Verifying access to the load balancer endpoint...");
        boolean wasSuccessul =
loadBalancer.verifyLoadBalancerEndpoint(elbDnsName);
        if (!wasSuccessul) {
            System.out.println("Couldn't connect to the load balancer, verifying
that the port is open...");
            CloseableHttpClient httpClient = HttpClients.createDefault();

            // Create an HTTP GET request to "http://checkip.amazonaws.com"
            HttpGet httpGet = new HttpGet("http://checkip.amazonaws.com");
            try {
                // Execute the request and get the response
                HttpResponse response = httpClient.execute(httpGet);

                // Read the response content.
                String ipAddress =
IOUtils.toString(response.getEntity().getContent(),
StandardCharsets.UTF_8).trim();

                // Print the public IP address.
                System.out.println("Public IP Address: " + ipAddress);
                GroupInfo groupInfo = autoScaler.verifyInboundPort(vpcId, port,
ipAddress);
                if (!groupInfo.isPortOpen()) {
                    System.out.println("""
                            For this example to work, the default security group
for your default VPC must
                            allow access from this computer. You can either add
it automatically from this
```

```
                            example or add it yourself using the AWS Management
Console.
                            """);

                    System.out.println(
                            "Do you want to add a rule to security group " +
groupInfo.getGroupName() + " to allow");
                    System.out.println("inbound traffic on port " + port + " from
your computer's IP address (y/n) ");
                    String ans = in.nextLine();
                    if ("y".equalsIgnoreCase(ans)) {
                        autoScaler.openInboundPort(groupInfo.getGroupName(),
String.valueOf(port), ipAddress);
                        System.out.println("Security group rule added.");
                    } else {
                        System.out.println("No security group rule added.");
                    }
                }

            } catch (AutoScalingException e) {
                e.printStackTrace();
            }
        } else if (wasSuccessul) {
            System.out.println("Your load balancer is ready. You can access it by
browsing to:");
            System.out.println("\t http://" + elbDnsName);
        } else {
            System.out.println("Couldn't get a successful response from the load
balancer endpoint. Troubleshoot by");
            System.out.println("manually verifying that your VPC and security
group are configured correctly and that");
            System.out.println("you can successfully make a GET request to the
load balancer.");
        }

        System.out.println("Press Enter when you're ready to continue with the
demo.");
        in.nextLine();
    }

    // A method that controls the demo part of the Java program.
    public static void demo(LoadBalancer loadBalancer) throws IOException,
InterruptedException {
        ParameterHelper paramHelper = new ParameterHelper();
```

```
        System.out.println("Read the ssm_only_policy.json file");
        String ssmOnlyPolicy = readFileAsString(ssmJSON);

        System.out.println("Resetting parameters to starting values for demo.");
        paramHelper.reset();

        System.out.println(
                """
                        This part of the demonstration shows how to toggle
different parts of the system
                        to create situations where the web service fails, and
shows how using a resilient
                        architecture can keep the web service running in spite
of these failures.

                        At the start, the load balancer endpoint returns
recommendations and reports that all targets are healthy.
                       """);
        demoChoices(loadBalancer);

        System.out.println(
                """
                        The web service running on the EC2 instances gets
recommendations by querying a DynamoDB table.
                        The table name is contained in a Systems Manager
parameter named self.param_helper.table.
                        To simulate a failure of the recommendation service,
let's set this parameter to name a non-existent table.
                       """);
        paramHelper.put(paramHelper.tableName, "this-is-not-a-table");

        System.out.println(
                """
                        \nNow, sending a GET request to the load balancer
endpoint returns a failure code. But, the service reports as
                        healthy to the load balancer because shallow health
checks don't check for failure of the recommendation service.
                       """);
        demoChoices(loadBalancer);

        System.out.println(
                """
                       Instead of failing when the recommendation service fails,
the web service can return a static response.
```

```
                        While this is not a perfect solution, it presents the
customer with a somewhat better experience than failure.
                        """);
        paramHelper.put(paramHelper.failureResponse, "static");

        System.out.println("""
                Now, sending a GET request to the load balancer endpoint returns
a static response.
                The service still reports as healthy because health checks are
still shallow.
                """);
        demoChoices(loadBalancer);

        System.out.println("Let's reinstate the recommendation service.");
        paramHelper.put(paramHelper.tableName, paramHelper.dyntable);

        System.out.println("""
                Let's also substitute bad credentials for one of the instances in
the target group so that it can't
                access the DynamoDB recommendation table. We will get an instance
id value.
                """);

        LaunchTemplateCreator templateCreator = new LaunchTemplateCreator();
        AutoScaler autoScaler = new AutoScaler();

        // Create a new instance profile based on badCredsProfileName.
        templateCreator.createInstanceProfile(policyFile, policyName,
badCredsProfileName, roleName);
        String badInstanceId = autoScaler.getBadInstance(autoScalingGroupName);
        System.out.println("The bad instance id values used for this demo is " +
badInstanceId);

        String profileAssociationId =
autoScaler.getInstanceProfile(badInstanceId);
        System.out.println("The association Id value is " +
profileAssociationId);
        System.out.println("Replacing the profile for instance " + badInstanceId
                + " with a profile that contains bad credentials");
        autoScaler.replaceInstanceProfile(badInstanceId, badCredsProfileName,
profileAssociationId);

        System.out.println(
                """
```

```
                        Now, sending a GET request to the load balancer endpoint
returns either a recommendation or a static response,
                        depending on which instance is selected by the load
balancer.
                        """);

        demoChoices(loadBalancer);

        System.out.println("""
                Let's implement a deep health check. For this demo, a deep health
check tests whether
                the web service can access the DynamoDB table that it depends on
for recommendations. Note that
                the deep health check is only for ELB routing and not for Auto
Scaling instance health.
                This kind of deep health check is not recommended for Auto
Scaling instance health, because it
                risks accidental termination of all instances in the Auto Scaling
group when a dependent service fails.
                """);

        System.out.println("""
                By implementing deep health checks, the load balancer can detect
when one of the instances is failing
                and take that instance out of rotation.
                """);

        paramHelper.put(paramHelper.healthCheck, "deep");

        System.out.println("""
                Now, checking target health indicates that the instance with bad
credentials
                is unhealthy. Note that it might take a minute or two for the
load balancer to detect the unhealthy
                instance. Sending a GET request to the load balancer endpoint
always returns a recommendation, because
                the load balancer takes unhealthy instances out of its rotation.
                """);

        demoChoices(loadBalancer);

        System.out.println(
                """
```

```
                        Because the instances in this demo are controlled by an
auto scaler, the simplest way to fix an unhealthy
                        instance is to terminate it and let the auto scaler start
a new instance to replace it.
                        """);
        autoScaler.terminateInstance(badInstanceId);

        System.out.println("""
                Even while the instance is terminating and the new instance is
starting, sending a GET
                request to the web service continues to get a successful
recommendation response because
                the load balancer routes requests to the healthy instances. After
the replacement instance
                starts and reports as healthy, it is included in the load
balancing rotation.
                Note that terminating and replacing an instance typically takes
several minutes, during which time you
                can see the changing health check status until the new instance
is running and healthy.
                """);

        demoChoices(loadBalancer);
        System.out.println(
                "If the recommendation service fails now, deep health checks mean
all instances report as unhealthy.");
        paramHelper.put(paramHelper.tableName, "this-is-not-a-table");

        demoChoices(loadBalancer);
        paramHelper.reset();
    }

    public static void demoChoices(LoadBalancer loadBalancer) throws IOException,
InterruptedException {
        String[] actions = {
                "Send a GET request to the load balancer endpoint.",
                "Check the health of load balancer targets.",
                "Go to the next part of the demo."
        };
        Scanner scanner = new Scanner(System.in);

        while (true) {
            System.out.println("-".repeat(88));
```

```
            System.out.println("See the current state of the service by selecting
one of the following choices:");
            for (int i = 0; i < actions.length; i++) {
                System.out.println(i + ": " + actions[i]);
            }

            try {
                System.out.print("\nWhich action would you like to take? ");
                int choice = scanner.nextInt();
                System.out.println("-".repeat(88));

                switch (choice) {
                    case 0 -> {
                        System.out.println("Request:\n");
                        System.out.println("GET http://" +
loadBalancer.getEndpoint(lbName));
                        CloseableHttpClient httpClient =
HttpClients.createDefault();

                        // Create an HTTP GET request to the ELB.
                        HttpGet httpGet = new HttpGet("http://" +
loadBalancer.getEndpoint(lbName));

                        // Execute the request and get the response.
                        HttpResponse response = httpClient.execute(httpGet);
                        int statusCode =
response.getStatusLine().getStatusCode();
                        System.out.println("HTTP Status Code: " + statusCode);

                        // Display the JSON response
                        BufferedReader reader = new BufferedReader(
                                new
InputStreamReader(response.getEntity().getContent()));
                        StringBuilder jsonResponse = new StringBuilder();
                        String line;
                        while ((line = reader.readLine()) != null) {
                            jsonResponse.append(line);
                        }
                        reader.close();

                        // Print the formatted JSON response.
                        System.out.println("Full Response:\n");
                        System.out.println(jsonResponse.toString());
```

```
                        // Close the HTTP client.
                        httpClient.close();

                    }
                    case 1 -> {
                        System.out.println("\nChecking the health of load
 balancer targets:\n");
                        List<TargetHealthDescription> health =
 loadBalancer.checkTargetHealth(targetGroupName);
                        for (TargetHealthDescription target : health) {
                            System.out.printf("\tTarget %s on port %d is %s%n",
 target.target().id(),
                                    target.target().port(),
 target.targetHealth().stateAsString());
                        }
                        System.out.println("""
                                Note that it can take a minute or two for the
 health check to update
                                after changes are made.
                                """);
                    }
                    case 2 -> {
                        System.out.println("\nOkay, let's move on.");
                        System.out.println("-".repeat(88));
                        return; // Exit the method when choice is 2
                    }
                    default -> System.out.println("You must choose a value
 between 0-2. Please select again.");
                }

            } catch (java.util.InputMismatchException e) {
                System.out.println("Invalid input. Please select again.");
                scanner.nextLine(); // Clear the input buffer.
            }
        }
    }

    public static String readFileAsString(String filePath) throws IOException {
        byte[] bytes = Files.readAllBytes(Paths.get(filePath));
        return new String(bytes);
    }
}
```

Crie uma classe que envolva ações do Auto Scaling e do Amazon EC2.

```
public class AutoScaler {

    private static Ec2Client ec2Client;
    private static AutoScalingClient autoScalingClient;
    private static IamClient iamClient;

    private static SsmClient ssmClient;

    private IamClient getIAMClient() {
        if (iamClient == null) {
            iamClient = IamClient.builder()
                    .region(Region.US_EAST_1)
                    .build();
        }
        return iamClient;
    }

    private SsmClient getSSMClient() {
        if (ssmClient == null) {
            ssmClient = SsmClient.builder()
                    .region(Region.US_EAST_1)
                    .build();
        }
        return ssmClient;
    }

    private Ec2Client getEc2Client() {
        if (ec2Client == null) {
            ec2Client = Ec2Client.builder()
                    .region(Region.US_EAST_1)
                    .build();
        }
        return ec2Client;
    }

    private AutoScalingClient getAutoScalingClient() {
        if (autoScalingClient == null) {
            autoScalingClient = AutoScalingClient.builder()
                    .region(Region.US_EAST_1)
                    .build();
        }
        return autoScalingClient;
```

```
   }

   /**
    * Terminates and instances in an EC2 Auto Scaling group. After an instance
is
    * terminated, it can no longer be accessed.
    */
   public void terminateInstance(String instanceId) {
       TerminateInstanceInAutoScalingGroupRequest terminateInstanceIRequest =
TerminateInstanceInAutoScalingGroupRequest
               .builder()
               .instanceId(instanceId)
               .shouldDecrementDesiredCapacity(false)
               .build();


getAutoScalingClient().terminateInstanceInAutoScalingGroup(terminateInstanceIRequest);
       System.out.format("Terminated instance %s.", instanceId);
   }

   /**
    * Replaces the profile associated with a running instance. After the profile
is
    * replaced, the instance is rebooted to ensure that it uses the new profile.
    * When
    * the instance is ready, Systems Manager is used to restart the Python web
    * server.
    */
   public void replaceInstanceProfile(String instanceId, String
newInstanceProfileName, String profileAssociationId)
           throws InterruptedException {
       // Create an IAM instance profile specification.
       software.amazon.awssdk.services.ec2.model.IamInstanceProfileSpecification
iamInstanceProfile =
software.amazon.awssdk.services.ec2.model.IamInstanceProfileSpecification
               .builder()
               .name(newInstanceProfileName) // Make sure
'newInstanceProfileName' is a valid IAM Instance Profile
                                             // name.
               .build();

       // Replace the IAM instance profile association for the EC2 instance.
       ReplaceIamInstanceProfileAssociationRequest replaceRequest =
ReplaceIamInstanceProfileAssociationRequest
```

```
                .builder()
                .iamInstanceProfile(iamInstanceProfile)
                .associationId(profileAssociationId) // Make sure
'profileAssociationId' is a valid association ID.
                .build();

        try {
            getEc2Client().replaceIamInstanceProfileAssociation(replaceRequest);
            // Handle the response as needed.
        } catch (Ec2Exception e) {
            // Handle exceptions, log, or report the error.
            System.err.println("Error: " + e.getMessage());
        }
        System.out.format("Replaced instance profile for association %s with
profile %s.", profileAssociationId,
                newInstanceProfileName);
        TimeUnit.SECONDS.sleep(15);
        boolean instReady = false;
        int tries = 0;

        // Reboot after 60 seconds
        while (!instReady) {
            if (tries % 6 == 0) {
                getEc2Client().rebootInstances(RebootInstancesRequest.builder()
                        .instanceIds(instanceId)
                        .build());
                System.out.println("Rebooting instance " + instanceId + " and
waiting for it to be ready.");
            }
            tries++;
            try {
                TimeUnit.SECONDS.sleep(10);
            } catch (InterruptedException e) {
                e.printStackTrace();
            }

            DescribeInstanceInformationResponse informationResponse =
getSSMClient().describeInstanceInformation();
            List<InstanceInformation> instanceInformationList =
informationResponse.instanceInformationList();
            for (InstanceInformation info : instanceInformationList) {
                if (info.instanceId().equals(instanceId)) {
                    instReady = true;
                    break;
```

```
                }
            }
        }

        SendCommandRequest sendCommandRequest = SendCommandRequest.builder()
                .instanceIds(instanceId)
                .documentName("AWS-RunShellScript")
                .parameters(Collections.singletonMap("commands",
                        Collections.singletonList("cd / && sudo python3 server.py
80")))
                .build();

        getSSMClient().sendCommand(sendCommandRequest);
        System.out.println("Restarted the Python web server on instance " +
instanceId + ".");
    }

    public void openInboundPort(String secGroupId, String port, String ipAddress)
{
        AuthorizeSecurityGroupIngressRequest ingressRequest =
AuthorizeSecurityGroupIngressRequest.builder()
                .groupName(secGroupId)
                .cidrIp(ipAddress)
                .fromPort(Integer.parseInt(port))
                .build();

        getEc2Client().authorizeSecurityGroupIngress(ingressRequest);
        System.out.format("Authorized ingress to %s on port %s from %s.",
secGroupId, port, ipAddress);
    }

    /**
     * Detaches a role from an instance profile, detaches policies from the role,
     * and deletes all the resources.
     */
    public void deleteInstanceProfile(String roleName, String profileName) {
        try {
            software.amazon.awssdk.services.iam.model.GetInstanceProfileRequest
getInstanceProfileRequest =
software.amazon.awssdk.services.iam.model.GetInstanceProfileRequest
                    .builder()
                    .instanceProfileName(profileName)
                    .build();
```

```
            GetInstanceProfileResponse response =
getIAMClient().getInstanceProfile(getInstanceProfileRequest);
            String name = response.instanceProfile().instanceProfileName();
            System.out.println(name);

            RemoveRoleFromInstanceProfileRequest profileRequest =
RemoveRoleFromInstanceProfileRequest.builder()
                    .instanceProfileName(profileName)
                    .roleName(roleName)
                    .build();

            getIAMClient().removeRoleFromInstanceProfile(profileRequest);
            DeleteInstanceProfileRequest deleteInstanceProfileRequest =
DeleteInstanceProfileRequest.builder()
                    .instanceProfileName(profileName)
                    .build();

            getIAMClient().deleteInstanceProfile(deleteInstanceProfileRequest);
            System.out.println("Deleted instance profile " + profileName);

            DeleteRoleRequest deleteRoleRequest = DeleteRoleRequest.builder()
                    .roleName(roleName)
                    .build();

            // List attached role policies.
            ListAttachedRolePoliciesResponse rolesResponse = getIAMClient()
                    .listAttachedRolePolicies(role -> role.roleName(roleName));
            List<AttachedPolicy> attachedPolicies =
rolesResponse.attachedPolicies();
            for (AttachedPolicy attachedPolicy : attachedPolicies) {
                DetachRolePolicyRequest request =
DetachRolePolicyRequest.builder()
                        .roleName(roleName)
                        .policyArn(attachedPolicy.policyArn())
                        .build();

                getIAMClient().detachRolePolicy(request);
                System.out.println("Detached and deleted policy " +
attachedPolicy.policyName());
            }

            getIAMClient().deleteRole(deleteRoleRequest);
            System.out.println("Instance profile and role deleted.");
```

```
        } catch (IamException e) {
            System.err.println(e.getMessage());
            System.exit(1);
        }
    }

    public void deleteTemplate(String templateName) {
        getEc2Client().deleteLaunchTemplate(name ->
name.launchTemplateName(templateName));
        System.out.format(templateName + " was deleted.");
    }

    public void deleteAutoScaleGroup(String groupName) {
        DeleteAutoScalingGroupRequest deleteAutoScalingGroupRequest =
DeleteAutoScalingGroupRequest.builder()
                .autoScalingGroupName(groupName)
                .forceDelete(true)
                .build();


getAutoScalingClient().deleteAutoScalingGroup(deleteAutoScalingGroupRequest);
        System.out.println(groupName + " was deleted.");
    }

    /*
     * Verify the default security group of the specified VPC allows ingress from
     * this
     * computer. This can be done by allowing ingress from this computer's IP
     * address. In some situations, such as connecting from a corporate network,
you
     * must instead specify a prefix list ID. You can also temporarily open the
port
     * to
     * any IP address while running this example. If you do, be sure to remove
     * public
     * access when you're done.
     *
     */
    public GroupInfo verifyInboundPort(String VPC, int port, String ipAddress) {
        boolean portIsOpen = false;
        GroupInfo groupInfo = new GroupInfo();
        try {
            Filter filter = Filter.builder()
                    .name("group-name")
```

```
                    .values("default")
                    .build();

            Filter filter1 = Filter.builder()
                    .name("vpc-id")
                    .values(VPC)
                    .build();

            DescribeSecurityGroupsRequest securityGroupsRequest =
DescribeSecurityGroupsRequest.builder()
                    .filters(filter, filter1)
                    .build();

            DescribeSecurityGroupsResponse securityGroupsResponse =
getEc2Client()
                    .describeSecurityGroups(securityGroupsRequest);
            String securityGroup =
securityGroupsResponse.securityGroups().get(0).groupName();
            groupInfo.setGroupName(securityGroup);

            for (SecurityGroup secGroup :
securityGroupsResponse.securityGroups()) {
                System.out.println("Found security group: " +
secGroup.groupId());

                for (IpPermission ipPermission : secGroup.ipPermissions()) {
                    if (ipPermission.fromPort() == port) {
                        System.out.println("Found inbound rule: " +
ipPermission);
                        for (IpRange ipRange : ipPermission.ipRanges()) {
                            String cidrIp = ipRange.cidrIp();
                            if (cidrIp.startsWith(ipAddress) ||
cidrIp.equals("0.0.0.0/0")) {
                                System.out.println(cidrIp + " is applicable");
                                portIsOpen = true;
                            }
                        }

                        if (!ipPermission.prefixListIds().isEmpty()) {
                            System.out.println("Prefix lList is applicable");
                            portIsOpen = true;
                        }

                        if (!portIsOpen) {
```

```
                            System.out
                                    .println("The inbound rule does not appear to
be open to either this computer's IP,"
                                            + " all IP addresses (0.0.0.0/0), or
to a prefix list ID.");
                        } else {
                            break;
                        }
                    }
                }
            }

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
        }

        groupInfo.setPortOpen(portIsOpen);
        return groupInfo;
    }

    /*
     * Attaches an Elastic Load Balancing (ELB) target group to this EC2 Auto
     * Scaling group.
     * The target group specifies how the load balancer forward requests to the
     * instances
     * in the group.
     */
    public void attachLoadBalancerTargetGroup(String asGroupName, String
targetGroupARN) {
        try {
            AttachLoadBalancerTargetGroupsRequest targetGroupsRequest =
AttachLoadBalancerTargetGroupsRequest.builder()
                    .autoScalingGroupName(asGroupName)
                    .targetGroupARNs(targetGroupARN)
                    .build();


getAutoScalingClient().attachLoadBalancerTargetGroups(targetGroupsRequest);
            System.out.println("Attached load balancer to " + asGroupName);

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
```

```
   }

   // Creates an EC2 Auto Scaling group with the specified size.
   public String[] createGroup(int groupSize, String templateName, String
autoScalingGroupName) {

       // Get availability zones.

software.amazon.awssdk.services.ec2.model.DescribeAvailabilityZonesRequest
zonesRequest =
software.amazon.awssdk.services.ec2.model.DescribeAvailabilityZonesRequest
               .builder()
               .build();

       DescribeAvailabilityZonesResponse zonesResponse =
getEc2Client().describeAvailabilityZones(zonesRequest);
       List<String> availabilityZoneNames =
zonesResponse.availabilityZones().stream()

.map(software.amazon.awssdk.services.ec2.model.AvailabilityZone::zoneName)
               .collect(Collectors.toList());

       String availabilityZones = String.join(",", availabilityZoneNames);
       LaunchTemplateSpecification specification =
LaunchTemplateSpecification.builder()
               .launchTemplateName(templateName)
               .version("$Default")
               .build();

       String[] zones = availabilityZones.split(",");
       CreateAutoScalingGroupRequest groupRequest =
CreateAutoScalingGroupRequest.builder()
               .launchTemplate(specification)
               .availabilityZones(zones)
               .maxSize(groupSize)
               .minSize(groupSize)
               .autoScalingGroupName(autoScalingGroupName)
               .build();

       try {
           getAutoScalingClient().createAutoScalingGroup(groupRequest);

       } catch (AutoScalingException e) {
           System.err.println(e.awsErrorDetails().errorMessage());
```

```
            System.exit(1);
        }
        System.out.println("Created an EC2 Auto Scaling group named " +
autoScalingGroupName);
        return zones;
    }

    public String getDefaultVPC() {
        // Define the filter.
        Filter defaultFilter = Filter.builder()
                .name("is-default")
                .values("true")
                .build();

        software.amazon.awssdk.services.ec2.model.DescribeVpcsRequest request =
software.amazon.awssdk.services.ec2.model.DescribeVpcsRequest
                .builder()
                .filters(defaultFilter)
                .build();

        DescribeVpcsResponse response = getEc2Client().describeVpcs(request);
        return response.vpcs().get(0).vpcId();
    }

    // Gets the default subnets in a VPC for a specified list of Availability
Zones.
    public List<Subnet> getSubnets(String vpcId, String[] availabilityZones) {
        List<Subnet> subnets = null;
        Filter vpcFilter = Filter.builder()
                .name("vpc-id")
                .values(vpcId)
                .build();

        Filter azFilter = Filter.builder()
                .name("availability-zone")
                .values(availabilityZones)
                .build();

        Filter defaultForAZ = Filter.builder()
                .name("default-for-az")
                .values("true")
                .build();

        DescribeSubnetsRequest request = DescribeSubnetsRequest.builder()
```

```
                .filters(vpcFilter, azFilter, defaultForAZ)
                .build();

        DescribeSubnetsResponse response =
getEc2Client().describeSubnets(request);
        subnets = response.subnets();
        return subnets;
    }

    // Gets data about the instances in the EC2 Auto Scaling group.
    public String getBadInstance(String groupName) {
        DescribeAutoScalingGroupsRequest request =
DescribeAutoScalingGroupsRequest.builder()
                .autoScalingGroupNames(groupName)
                .build();

        DescribeAutoScalingGroupsResponse response =
getAutoScalingClient().describeAutoScalingGroups(request);
        AutoScalingGroup autoScalingGroup = response.autoScalingGroups().get(0);
        List<String> instanceIds = autoScalingGroup.instances().stream()
                .map(instance -> instance.instanceId())
                .collect(Collectors.toList());

        String[] instanceIdArray = instanceIds.toArray(new String[0]);
        for (String instanceId : instanceIdArray) {
            System.out.println("Instance ID: " + instanceId);
            return instanceId;
        }
        return "";
    }

    // Gets data about the profile associated with an instance.
    public String getInstanceProfile(String instanceId) {
        Filter filter = Filter.builder()
                .name("instance-id")
                .values(instanceId)
                .build();

        DescribeIamInstanceProfileAssociationsRequest associationsRequest =
DescribeIamInstanceProfileAssociationsRequest
                .builder()
                .filters(filter)
                .build();
```

```
        DescribeIamInstanceProfileAssociationsResponse response = getEc2Client()
                .describeIamInstanceProfileAssociations(associationsRequest);
        return response.iamInstanceProfileAssociations().get(0).associationId();
    }

    public void deleteRolesPolicies(String policyName, String roleName, String
 InstanceProfile) {
        ListPoliciesRequest listPoliciesRequest =
 ListPoliciesRequest.builder().build();
        ListPoliciesResponse listPoliciesResponse =
 getIAMClient().listPolicies(listPoliciesRequest);
        for (Policy policy : listPoliciesResponse.policies()) {
            if (policy.policyName().equals(policyName)) {
                // List the entities (users, groups, roles) that are attached to
 the policy.

 software.amazon.awssdk.services.iam.model.ListEntitiesForPolicyRequest
 listEntitiesRequest =
 software.amazon.awssdk.services.iam.model.ListEntitiesForPolicyRequest
                        .builder()
                        .policyArn(policy.arn())
                        .build();
                ListEntitiesForPolicyResponse listEntitiesResponse = iamClient
                        .listEntitiesForPolicy(listEntitiesRequest);
                if (!listEntitiesResponse.policyGroups().isEmpty() || !
listEntitiesResponse.policyUsers().isEmpty()
                        || !listEntitiesResponse.policyRoles().isEmpty()) {
                    // Detach the policy from any entities it is attached to.
                    DetachRolePolicyRequest detachPolicyRequest =
 DetachRolePolicyRequest.builder()
                            .policyArn(policy.arn())
                            .roleName(roleName) // Specify the name of the IAM
 role
                            .build();

                    getIAMClient().detachRolePolicy(detachPolicyRequest);
                    System.out.println("Policy detached from entities.");
                }

                // Now, you can delete the policy.
                DeletePolicyRequest deletePolicyRequest =
 DeletePolicyRequest.builder()
                        .policyArn(policy.arn())
                        .build();
```

```
                getIAMClient().deletePolicy(deletePolicyRequest);
                System.out.println("Policy deleted successfully.");
                break;
            }
        }

        // List the roles associated with the instance profile
        ListInstanceProfilesForRoleRequest listRolesRequest =
 ListInstanceProfilesForRoleRequest.builder()
                .roleName(roleName)
                .build();

        // Detach the roles from the instance profile
        ListInstanceProfilesForRoleResponse listRolesResponse =
 iamClient.listInstanceProfilesForRole(listRolesRequest);
        for (software.amazon.awssdk.services.iam.model.InstanceProfile profile :
 listRolesResponse.instanceProfiles()) {
            RemoveRoleFromInstanceProfileRequest removeRoleRequest =
 RemoveRoleFromInstanceProfileRequest.builder()
                    .instanceProfileName(InstanceProfile)
                    .roleName(roleName) // Remove the extra dot here
                    .build();

            getIAMClient().removeRoleFromInstanceProfile(removeRoleRequest);
            System.out.println("Role " + roleName + " removed from instance
 profile " + InstanceProfile);
        }

        // Delete the instance profile after removing all roles
        DeleteInstanceProfileRequest deleteInstanceProfileRequest =
 DeleteInstanceProfileRequest.builder()
                .instanceProfileName(InstanceProfile)
                .build();

        getIAMClient().deleteInstanceProfile(r ->
 r.instanceProfileName(InstanceProfile));
        System.out.println(InstanceProfile + " Deleted");
        System.out.println("All roles and policies are deleted.");
    }
}
```

Crie uma classe que envolva ações do Elastic Load Balancing.

```
public class LoadBalancer {
    public ElasticLoadBalancingV2Client elasticLoadBalancingV2Client;

    public ElasticLoadBalancingV2Client getLoadBalancerClient() {
        if (elasticLoadBalancingV2Client == null) {
            elasticLoadBalancingV2Client = ElasticLoadBalancingV2Client.builder()
                    .region(Region.US_EAST_1)
                    .build();
        }

        return elasticLoadBalancingV2Client;
    }

    // Checks the health of the instances in the target group.
    public List<TargetHealthDescription> checkTargetHealth(String
 targetGroupName) {
        DescribeTargetGroupsRequest targetGroupsRequest =
 DescribeTargetGroupsRequest.builder()
                .names(targetGroupName)
                .build();

        DescribeTargetGroupsResponse tgResponse =
 getLoadBalancerClient().describeTargetGroups(targetGroupsRequest);

        DescribeTargetHealthRequest healthRequest =
 DescribeTargetHealthRequest.builder()

 .targetGroupArn(tgResponse.targetGroups().get(0).targetGroupArn())
                .build();

        DescribeTargetHealthResponse healthResponse =
 getLoadBalancerClient().describeTargetHealth(healthRequest);
        return healthResponse.targetHealthDescriptions();
    }

    // Gets the HTTP endpoint of the load balancer.
    public String getEndpoint(String lbName) {
        DescribeLoadBalancersResponse res = getLoadBalancerClient()
                .describeLoadBalancers(describe -> describe.names(lbName));
        return res.loadBalancers().get(0).dnsName();
    }
```

```
   // Deletes a load balancer.
   public void deleteLoadBalancer(String lbName) {
       try {
           // Use a waiter to delete the Load Balancer.
           DescribeLoadBalancersResponse res = getLoadBalancerClient()
                   .describeLoadBalancers(describe -> describe.names(lbName));
           ElasticLoadBalancingV2Waiter loadBalancerWaiter =
getLoadBalancerClient().waiter();
           DescribeLoadBalancersRequest request =
DescribeLoadBalancersRequest.builder()

.loadBalancerArns(res.loadBalancers().get(0).loadBalancerArn())
                   .build();

           getLoadBalancerClient().deleteLoadBalancer(
                   builder ->
builder.loadBalancerArn(res.loadBalancers().get(0).loadBalancerArn()));
           WaiterResponse<DescribeLoadBalancersResponse> waiterResponse =
loadBalancerWaiter
                   .waitUntilLoadBalancersDeleted(request);
           waiterResponse.matched().response().ifPresent(System.out::println);

       } catch (ElasticLoadBalancingV2Exception e) {
           System.err.println(e.awsErrorDetails().errorMessage());
       }
       System.out.println(lbName + " was deleted.");
   }

   // Deletes the target group.
   public void deleteTargetGroup(String targetGroupName) {
       try {
           DescribeTargetGroupsResponse res = getLoadBalancerClient()
                   .describeTargetGroups(describe ->
describe.names(targetGroupName));
           getLoadBalancerClient()
                   .deleteTargetGroup(builder ->
builder.targetGroupArn(res.targetGroups().get(0).targetGroupArn()));
       } catch (ElasticLoadBalancingV2Exception e) {
           System.err.println(e.awsErrorDetails().errorMessage());
       }
       System.out.println(targetGroupName + " was deleted.");
   }
```

```
   // Verify this computer can successfully send a GET request to the load
balancer
   // endpoint.
   public boolean verifyLoadBalancerEndpoint(String elbDnsName) throws
IOException, InterruptedException {
       boolean success = false;
       int retries = 3;
       CloseableHttpClient httpClient = HttpClients.createDefault();

       // Create an HTTP GET request to the ELB.
       HttpGet httpGet = new HttpGet("http://" + elbDnsName);
       try {
           while ((!success) && (retries > 0)) {
               // Execute the request and get the response.
               HttpResponse response = httpClient.execute(httpGet);
               int statusCode = response.getStatusLine().getStatusCode();
               System.out.println("HTTP Status Code: " + statusCode);
               if (statusCode == 200) {
                   success = true;
               } else {
                   retries--;
                   System.out.println("Got connection error from load balancer
endpoint, retrying...");
                   TimeUnit.SECONDS.sleep(15);
               }
           }

       } catch (org.apache.http.conn.HttpHostConnectException e) {
           System.out.println(e.getMessage());
       }

       System.out.println("Status.." + success);
       return success;
   }

   /*
    * Creates an Elastic Load Balancing target group. The target group specifies
    * how
    * the load balancer forward requests to instances in the group and how
instance
    * health is checked.
    */
   public String createTargetGroup(String protocol, int port, String vpcId,
String targetGroupName) {
```

```
        CreateTargetGroupRequest targetGroupRequest =
CreateTargetGroupRequest.builder()
                .healthCheckPath("/healthcheck")
                .healthCheckTimeoutSeconds(5)
                .port(port)
                .vpcId(vpcId)
                .name(targetGroupName)
                .protocol(protocol)
                .build();

        CreateTargetGroupResponse targetGroupResponse =
getLoadBalancerClient().createTargetGroup(targetGroupRequest);
        String targetGroupArn =
targetGroupResponse.targetGroups().get(0).targetGroupArn();
        String targetGroup =
targetGroupResponse.targetGroups().get(0).targetGroupName();
        System.out.println("The " + targetGroup + " was created with ARN" +
targetGroupArn);
        return targetGroupArn;
    }

    /*
     * Creates an Elastic Load Balancing load balancer that uses the specified
     * subnets
     * and forwards requests to the specified target group.
     */
    public String createLoadBalancer(List<Subnet> subnetIds, String
targetGroupARN, String lbName, int port,
            String protocol) {
        try {
            List<String> subnetIdStrings = subnetIds.stream()
                    .map(Subnet::subnetId)
                    .collect(Collectors.toList());

            CreateLoadBalancerRequest balancerRequest =
CreateLoadBalancerRequest.builder()
                    .subnets(subnetIdStrings)
                    .name(lbName)
                    .scheme("internet-facing")
                    .build();

            // Create and wait for the load balancer to become available.
            CreateLoadBalancerResponse lsResponse =
getLoadBalancerClient().createLoadBalancer(balancerRequest);
```

```
            String lbARN = lsResponse.loadBalancers().get(0).loadBalancerArn();

            ElasticLoadBalancingV2Waiter loadBalancerWaiter =
getLoadBalancerClient().waiter();
            DescribeLoadBalancersRequest request =
DescribeLoadBalancersRequest.builder()
                    .loadBalancerArns(lbARN)
                    .build();

            System.out.println("Waiting for Load Balancer " + lbName + " to
become available.");
            WaiterResponse<DescribeLoadBalancersResponse> waiterResponse =
loadBalancerWaiter
                    .waitUntilLoadBalancerAvailable(request);
            waiterResponse.matched().response().ifPresent(System.out::println);
            System.out.println("Load Balancer " + lbName + " is available.");

            // Get the DNS name (endpoint) of the load balancer.
            String lbDNSName = lsResponse.loadBalancers().get(0).dnsName();
            System.out.println("*** Load Balancer DNS Name: " + lbDNSName);

            // Create a listener for the load balance.
            Action action = Action.builder()
                    .targetGroupArn(targetGroupARN)
                    .type("forward")
                    .build();

            CreateListenerRequest listenerRequest =
CreateListenerRequest.builder()

.loadBalancerArn(lsResponse.loadBalancers().get(0).loadBalancerArn())
                    .defaultActions(action)
                    .port(port)
                    .protocol(protocol)
                    .build();

            getLoadBalancerClient().createListener(listenerRequest);
            System.out.println("Created listener to forward traffic from load
balancer " + lbName + " to target group "
                    + targetGroupARN);

            // Return the load balancer DNS name.
            return lbDNSName;
```

```
        } catch (ElasticLoadBalancingV2Exception e) {
            e.printStackTrace();
        }
        return "";
    }
}
```

Crie uma classe que use o DynamoDB para simular um serviço de recomendação.

```
public class Database {

    private static DynamoDbClient dynamoDbClient;

    public static DynamoDbClient getDynamoDbClient() {
        if (dynamoDbClient == null) {
            dynamoDbClient = DynamoDbClient.builder()
                    .region(Region.US_EAST_1)
                    .build();
        }
        return dynamoDbClient;
    }

    // Checks to see if the Amazon DynamoDB table exists.
    private boolean doesTableExist(String tableName) {
        try {
            // Describe the table and catch any exceptions.
            DescribeTableRequest describeTableRequest =
 DescribeTableRequest.builder()
                    .tableName(tableName)
                    .build();

            getDynamoDbClient().describeTable(describeTableRequest);
            System.out.println("Table '" + tableName + "' exists.");
            return true;

        } catch (ResourceNotFoundException e) {
            System.out.println("Table '" + tableName + "' does not exist.");
        } catch (DynamoDbException e) {
            System.err.println("Error checking table existence: " +
 e.getMessage());
        }
        return false;
```

```
    }

    /*
     * Creates a DynamoDB table to use a recommendation service. The table has a
     * hash key named 'MediaType' that defines the type of media recommended,
such
     * as
     * Book or Movie, and a range key named 'ItemId' that, combined with the
     * MediaType,
     * forms a unique identifier for the recommended item.
     */
    public void createTable(String tableName, String fileName) throws IOException
{
        // First check to see if the table exists.
        boolean doesExist = doesTableExist(tableName);
        if (!doesExist) {
            DynamoDbWaiter dbWaiter = getDynamoDbClient().waiter();
            CreateTableRequest createTableRequest = CreateTableRequest.builder()
                    .tableName(tableName)
                    .attributeDefinitions(
                            AttributeDefinition.builder()
                                    .attributeName("MediaType")
                                    .attributeType(ScalarAttributeType.S)
                                    .build(),
                            AttributeDefinition.builder()
                                    .attributeName("ItemId")
                                    .attributeType(ScalarAttributeType.N)
                                    .build())
                    .keySchema(
                            KeySchemaElement.builder()
                                    .attributeName("MediaType")
                                    .keyType(KeyType.HASH)
                                    .build(),
                            KeySchemaElement.builder()
                                    .attributeName("ItemId")
                                    .keyType(KeyType.RANGE)
                                    .build())
                    .provisionedThroughput(
                            ProvisionedThroughput.builder()
                                    .readCapacityUnits(5L)
                                    .writeCapacityUnits(5L)
                                    .build())
                    .build();
```

```
            getDynamoDbClient().createTable(createTableRequest);
            System.out.println("Creating table " + tableName + "...");

            // Wait until the Amazon DynamoDB table is created.
            DescribeTableRequest tableRequest = DescribeTableRequest.builder()
                    .tableName(tableName)
                    .build();

            WaiterResponse<DescribeTableResponse> waiterResponse =
dbWaiter.waitUntilTableExists(tableRequest);
            waiterResponse.matched().response().ifPresent(System.out::println);
            System.out.println("Table " + tableName + " created.");

            // Add records to the table.
            populateTable(fileName, tableName);
        }
    }

    public void deleteTable(String tableName) {
        getDynamoDbClient().deleteTable(table -> table.tableName(tableName));
        System.out.println("Table " + tableName + " deleted.");
    }

    // Populates the table with data located in a JSON file using the DynamoDB
    // enhanced client.
    public void populateTable(String fileName, String tableName) throws
IOException {
        DynamoDbEnhancedClient enhancedClient = DynamoDbEnhancedClient.builder()
                .dynamoDbClient(getDynamoDbClient())
                .build();
        ObjectMapper objectMapper = new ObjectMapper();
        File jsonFile = new File(fileName);
        JsonNode rootNode = objectMapper.readTree(jsonFile);

        DynamoDbTable<Recommendation> mappedTable =
enhancedClient.table(tableName,
                TableSchema.fromBean(Recommendation.class));
        for (JsonNode currentNode : rootNode) {
            String mediaType = currentNode.path("MediaType").path("S").asText();
            int itemId = currentNode.path("ItemId").path("N").asInt();
            String title = currentNode.path("Title").path("S").asText();
            String creator = currentNode.path("Creator").path("S").asText();

            // Create a Recommendation object and set its properties.
```

```
            Recommendation rec = new Recommendation();
            rec.setMediaType(mediaType);
            rec.setItemId(itemId);
            rec.setTitle(title);
            rec.setCreator(creator);

            // Put the item into the DynamoDB table.
            mappedTable.putItem(rec); // Add the Recommendation to the list.
        }
        System.out.println("Added all records to the " + tableName);
    }
}
```

Crie uma classe que envolva as ações do Systems Manager.

```
public class ParameterHelper {

    String tableName = "doc-example-resilient-architecture-table";
    String dyntable = "doc-example-recommendation-service";
    String failureResponse = "doc-example-resilient-architecture-failure-
response";
    String healthCheck = "doc-example-resilient-architecture-health-check";

    public void reset() {
        put(dyntable, tableName);
        put(failureResponse, "none");
        put(healthCheck, "shallow");
    }

    public void put(String name, String value) {
        SsmClient ssmClient = SsmClient.builder()
                .region(Region.US_EAST_1)
                .build();

        PutParameterRequest parameterRequest = PutParameterRequest.builder()
                .name(name)
                .value(value)
                .overwrite(true)
                .type("String")
                .build();

        ssmClient.putParameter(parameterRequest);
```

```
        System.out.printf("Setting demo parameter %s to '%s'.", name, value);
    }
}
```

- Para obter detalhes da API, consulte os tópicos a seguir na Referência da API AWS SDK for Java 2.x .
  - AttachLoadBalancerTargetGroups
  - CreateAutoScalingGroup
  - CreateInstanceProfile
  - CreateLaunchTemplate
  - CreateListener
  - CreateLoadBalancer
  - CreateTargetGroup
  - DeleteAutoScalingGroup
  - DeleteInstanceProfile
  - DeleteLaunchTemplate
  - DeleteLoadBalancer
  - DeleteTargetGroup
  - DescribeAutoScalingGroups
  - DescribeAvailabilityZones
  - DescribeIamInstanceProfileAssociations
  - DescribeInstances
  - DescribeLoadBalancers
  - DescribeSubnets
  - DescribeTargetGroups
  - DescribeTargetHealth
  - DescribeVpcs
  - RebootInstances
  - ReplaceIamInstanceProfileAssociation
  - TerminateInstanceInAutoScalingGroup
  - UpdateAutoScalingGroup

JavaScript

SDK para JavaScript (v3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

Execute o cenário interativo em um prompt de comando.

```
#!/usr/bin/env node
// Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.
// SPDX-License-Identifier: Apache-2.0

import {
  Scenario,
  parseScenarioArgs,
} from "@aws-doc-sdk-examples/lib/scenario/index.js";

/**
 * The workflow steps are split into three stages:
 *   - deploy
 *   - demo
 *   - destroy
 *
 * Each of these stages has a corresponding file prefixed with steps-*.
 */
import { deploySteps } from "./steps-deploy.js";
import { demoSteps } from "./steps-demo.js";
import { destroySteps } from "./steps-destroy.js";

/**
 * The context is passed to every scenario. Scenario steps
 * will modify the context.
 */
const context = {};

/**
 * Three Scenarios are created for the workflow. A Scenario is an orchestration
 class
 * that simplifies running a series of steps.
```

```
 */
export const scenarios = {
  // Deploys all resources necessary for the workflow.
  deploy: new Scenario("Resilient Workflow - Deploy", deploySteps, context),
  // Demonstrates how a fragile web service can be made more resilient.
  demo: new Scenario("Resilient Workflow - Demo", demoSteps, context),
  // Destroys the resources created for the workflow.
  destroy: new Scenario("Resilient Workflow - Destroy", destroySteps, context),
};

// Call function if run directly
import { fileURLToPath } from "node:url";

if (process.argv[1] === fileURLToPath(import.meta.url)) {
  parseScenarioArgs(scenarios, {
    name: "Resilient Workflow",
    synopsis:
      "node index.js --scenario <deploy | demo | destroy> [-h|--help] [-y|--yes]
 [-v|--verbose]",
    description: "Deploy and interact with scalable EC2 instances.",
  });
}
```

Criar etapas para implantar todos os recursos.

```
import { join } from "node:path";
import { readFileSync, writeFileSync } from "node:fs";
import axios from "axios";

import {
  BatchWriteItemCommand,
  CreateTableCommand,
  DynamoDBClient,
  waitUntilTableExists,
} from "@aws-sdk/client-dynamodb";
import {
  EC2Client,
  CreateKeyPairCommand,
  CreateLaunchTemplateCommand,
  DescribeAvailabilityZonesCommand,
  DescribeVpcsCommand,
  DescribeSubnetsCommand,
```

```
  DescribeSecurityGroupsCommand,
  AuthorizeSecurityGroupIngressCommand,
} from "@aws-sdk/client-ec2";
import {
  IAMClient,
  CreatePolicyCommand,
  CreateRoleCommand,
  CreateInstanceProfileCommand,
  AddRoleToInstanceProfileCommand,
  AttachRolePolicyCommand,
  waitUntilInstanceProfileExists,
} from "@aws-sdk/client-iam";
import { SSMClient, GetParameterCommand } from "@aws-sdk/client-ssm";
import {
  CreateAutoScalingGroupCommand,
  AutoScalingClient,
  AttachLoadBalancerTargetGroupsCommand,
} from "@aws-sdk/client-auto-scaling";
import {
  CreateListenerCommand,
  CreateLoadBalancerCommand,
  CreateTargetGroupCommand,
  ElasticLoadBalancingV2Client,
  waitUntilLoadBalancerAvailable,
} from "@aws-sdk/client-elastic-load-balancing-v2";

import {
  ScenarioOutput,
  ScenarioInput,
  ScenarioAction,
} from "@aws-doc-sdk-examples/lib/scenario/index.js";
import { saveState } from "@aws-doc-sdk-examples/lib/scenario/steps-common.js";
import { retry } from "@aws-doc-sdk-examples/lib/utils/util-timers.js";

import { MESSAGES, NAMES, RESOURCES_PATH, ROOT } from "./constants.js";
import { initParamsSteps } from "./steps-reset-params.js";

/**
 * @type {import('@aws-doc-sdk-examples/lib/scenario.js').Step[]}
 */
export const deploySteps = [
  new ScenarioOutput("introduction", MESSAGES.introduction, { header: true }),
  new ScenarioInput("confirmDeployment", MESSAGES.confirmDeployment, {
    type: "confirm",
```

```
  }),
  new ScenarioAction(
    "handleConfirmDeployment",
    (c) => c.confirmDeployment === false && process.exit(),
  ),
  new ScenarioOutput(
    "creatingTable",
    MESSAGES.creatingTable.replace("${TABLE_NAME}", NAMES.tableName),
  ),
  new ScenarioAction("createTable", async () => {
    const client = new DynamoDBClient({});
    await client.send(
      new CreateTableCommand({
        TableName: NAMES.tableName,
        ProvisionedThroughput: {
          ReadCapacityUnits: 5,
          WriteCapacityUnits: 5,
        },
        AttributeDefinitions: [
          {
            AttributeName: "MediaType",
            AttributeType: "S",
          },
          {
            AttributeName: "ItemId",
            AttributeType: "N",
          },
        ],
        KeySchema: [
          {
            AttributeName: "MediaType",
            KeyType: "HASH",
          },
          {
            AttributeName: "ItemId",
            KeyType: "RANGE",
          },
        ],
      }),
    );
    await waitUntilTableExists({ client }, { TableName: NAMES.tableName });
  }),
  new ScenarioOutput(
    "createdTable",
```

```
    MESSAGES.createdTable.replace("${TABLE_NAME}", NAMES.tableName),
  ),
  new ScenarioOutput(
    "populatingTable",
    MESSAGES.populatingTable.replace("${TABLE_NAME}", NAMES.tableName),
  ),
  new ScenarioAction("populateTable", () => {
    const client = new DynamoDBClient({});
    /**
     * @type {{ default: import("@aws-sdk/client-dynamodb").PutRequest['Item']
[] }}
     */
    const recommendations = JSON.parse(
      readFileSync(join(RESOURCES_PATH, "recommendations.json")),
    );

    return client.send(
      new BatchWriteItemCommand({
        RequestItems: {
          [NAMES.tableName]: recommendations.map((item) => ({
            PutRequest: { Item: item },
          })),
        },
      }),
    );
  }),
  new ScenarioOutput(
    "populatedTable",
    MESSAGES.populatedTable.replace("${TABLE_NAME}", NAMES.tableName),
  ),
  new ScenarioOutput(
    "creatingKeyPair",
    MESSAGES.creatingKeyPair.replace("${KEY_PAIR_NAME}", NAMES.keyPairName),
  ),
  new ScenarioAction("createKeyPair", async () => {
    const client = new EC2Client({});
    const { KeyMaterial } = await client.send(
      new CreateKeyPairCommand({
        KeyName: NAMES.keyPairName,
      }),
    );

    writeFileSync(`${NAMES.keyPairName}.pem`, KeyMaterial, { mode: 0o600 });
  }),
```

```
  new ScenarioOutput(
    "createdKeyPair",
    MESSAGES.createdKeyPair.replace("${KEY_PAIR_NAME}", NAMES.keyPairName),
  ),
  new ScenarioOutput(
    "creatingInstancePolicy",
    MESSAGES.creatingInstancePolicy.replace(
      "${INSTANCE_POLICY_NAME}",
      NAMES.instancePolicyName,
    ),
  ),
  new ScenarioAction("createInstancePolicy", async (state) => {
    const client = new IAMClient({});
    const {
      Policy: { Arn },
    } = await client.send(
      new CreatePolicyCommand({
        PolicyName: NAMES.instancePolicyName,
        PolicyDocument: readFileSync(
          join(RESOURCES_PATH, "instance_policy.json"),
        ),
      }),
    );
    state.instancePolicyArn = Arn;
  }),
  new ScenarioOutput("createdInstancePolicy", (state) =>
    MESSAGES.createdInstancePolicy
      .replace("${INSTANCE_POLICY_NAME}", NAMES.instancePolicyName)
      .replace("${INSTANCE_POLICY_ARN}", state.instancePolicyArn),
  ),
  new ScenarioOutput(
    "creatingInstanceRole",
    MESSAGES.creatingInstanceRole.replace(
      "${INSTANCE_ROLE_NAME}",
      NAMES.instanceRoleName,
    ),
  ),
  new ScenarioAction("createInstanceRole", () => {
    const client = new IAMClient({});
    return client.send(
      new CreateRoleCommand({
        RoleName: NAMES.instanceRoleName,
        AssumeRolePolicyDocument: readFileSync(
          join(ROOT, "assume-role-policy.json"),
```

```
      ),
    }),
  );
}),
new ScenarioOutput(
  "createdInstanceRole",
  MESSAGES.createdInstanceRole.replace(
    "${INSTANCE_ROLE_NAME}",
    NAMES.instanceRoleName,
  ),
),
new ScenarioOutput(
  "attachingPolicyToRole",
  MESSAGES.attachingPolicyToRole
    .replace("${INSTANCE_ROLE_NAME}", NAMES.instanceRoleName)
    .replace("${INSTANCE_POLICY_NAME}", NAMES.instancePolicyName),
),
new ScenarioAction("attachPolicyToRole", async (state) => {
  const client = new IAMClient({});
  await client.send(
    new AttachRolePolicyCommand({
      RoleName: NAMES.instanceRoleName,
      PolicyArn: state.instancePolicyArn,
    }),
  );
}),
new ScenarioOutput(
  "attachedPolicyToRole",
  MESSAGES.attachedPolicyToRole
    .replace("${INSTANCE_POLICY_NAME}", NAMES.instancePolicyName)
    .replace("${INSTANCE_ROLE_NAME}", NAMES.instanceRoleName),
),
new ScenarioOutput(
  "creatingInstanceProfile",
  MESSAGES.creatingInstanceProfile.replace(
    "${INSTANCE_PROFILE_NAME}",
    NAMES.instanceProfileName,
  ),
),
new ScenarioAction("createInstanceProfile", async (state) => {
  const client = new IAMClient({});
  const {
    InstanceProfile: { Arn },
  } = await client.send(
```

```
      new CreateInstanceProfileCommand({
        InstanceProfileName: NAMES.instanceProfileName,
      }),
    );
    state.instanceProfileArn = Arn;

    await waitUntilInstanceProfileExists(
      { client },
      { InstanceProfileName: NAMES.instanceProfileName },
    );
  }),
  new ScenarioOutput("createdInstanceProfile", (state) =>
    MESSAGES.createdInstanceProfile
      .replace("${INSTANCE_PROFILE_NAME}", NAMES.instanceProfileName)
      .replace("${INSTANCE_PROFILE_ARN}", state.instanceProfileArn),
  ),
  new ScenarioOutput(
    "addingRoleToInstanceProfile",
    MESSAGES.addingRoleToInstanceProfile
      .replace("${INSTANCE_PROFILE_NAME}", NAMES.instanceProfileName)
      .replace("${INSTANCE_ROLE_NAME}", NAMES.instanceRoleName),
  ),
  new ScenarioAction("addRoleToInstanceProfile", () => {
    const client = new IAMClient({});
    return client.send(
      new AddRoleToInstanceProfileCommand({
        RoleName: NAMES.instanceRoleName,
        InstanceProfileName: NAMES.instanceProfileName,
      }),
    );
  }),
  new ScenarioOutput(
    "addedRoleToInstanceProfile",
    MESSAGES.addedRoleToInstanceProfile
      .replace("${INSTANCE_PROFILE_NAME}", NAMES.instanceProfileName)
      .replace("${INSTANCE_ROLE_NAME}", NAMES.instanceRoleName),
  ),
  ...initParamsSteps,
  new ScenarioOutput("creatingLaunchTemplate", MESSAGES.creatingLaunchTemplate),
  new ScenarioAction("createLaunchTemplate", async () => {
    const ssmClient = new SSMClient({});
    const { Parameter } = await ssmClient.send(
      new GetParameterCommand({
        Name: "/aws/service/ami-amazon-linux-latest/amzn2-ami-hvm-x86_64-gp2",
```

```
      }),
    );
    const ec2Client = new EC2Client({});
    await ec2Client.send(
      new CreateLaunchTemplateCommand({
        LaunchTemplateName: NAMES.launchTemplateName,
        LaunchTemplateData: {
          InstanceType: "t3.micro",
          ImageId: Parameter.Value,
          IamInstanceProfile: { Name: NAMES.instanceProfileName },
          UserData: readFileSync(
            join(RESOURCES_PATH, "server_startup_script.sh"),
          ).toString("base64"),
          KeyName: NAMES.keyPairName,
        },
      }),
    );
  }),
  new ScenarioOutput(
    "createdLaunchTemplate",
    MESSAGES.createdLaunchTemplate.replace(
      "${LAUNCH_TEMPLATE_NAME}",
      NAMES.launchTemplateName,
    ),
  ),
  new ScenarioOutput(
    "creatingAutoScalingGroup",
    MESSAGES.creatingAutoScalingGroup.replace(
      "${AUTO_SCALING_GROUP_NAME}",
      NAMES.autoScalingGroupName,
    ),
  ),
  new ScenarioAction("createAutoScalingGroup", async (state) => {
    const ec2Client = new EC2Client({});
    const { AvailabilityZones } = await ec2Client.send(
      new DescribeAvailabilityZonesCommand({}),
    );
    state.availabilityZoneNames = AvailabilityZones.map((az) => az.ZoneName);
    const autoScalingClient = new AutoScalingClient({});
    await retry({ intervalInMs: 1000, maxRetries: 30 }, () =>
      autoScalingClient.send(
        new CreateAutoScalingGroupCommand({
          AvailabilityZones: state.availabilityZoneNames,
          AutoScalingGroupName: NAMES.autoScalingGroupName,
```

```
        LaunchTemplate: {
          LaunchTemplateName: NAMES.launchTemplateName,
          Version: "$Default",
        },
        MinSize: 3,
        MaxSize: 3,
      }),
    ),
  );
}),
new ScenarioOutput(
  "createdAutoScalingGroup",
  /**
   * @param {{ availabilityZoneNames: string[] }} state
   */
  (state) =>
    MESSAGES.createdAutoScalingGroup
      .replace("${AUTO_SCALING_GROUP_NAME}", NAMES.autoScalingGroupName)
      .replace(
        "${AVAILABILITY_ZONE_NAMES}",
        state.availabilityZoneNames.join(", "),
      ),
),
new ScenarioInput("confirmContinue", MESSAGES.confirmContinue, {
  type: "confirm",
}),
new ScenarioOutput("loadBalancer", MESSAGES.loadBalancer),
new ScenarioOutput("gettingVpc", MESSAGES.gettingVpc),
new ScenarioAction("getVpc", async (state) => {
  const client = new EC2Client({});
  const { Vpcs } = await client.send(
    new DescribeVpcsCommand({
      Filters: [{ Name: "is-default", Values: ["true"] }],
    }),
  );
  state.defaultVpc = Vpcs[0].VpcId;
}),
new ScenarioOutput("gotVpc", (state) =>
  MESSAGES.gotVpc.replace("${VPC_ID}", state.defaultVpc),
),
new ScenarioOutput("gettingSubnets", MESSAGES.gettingSubnets),
new ScenarioAction("getSubnets", async (state) => {
  const client = new EC2Client({});
  const { Subnets } = await client.send(
```

```
    new DescribeSubnetsCommand({
      Filters: [
        { Name: "vpc-id", Values: [state.defaultVpc] },
        { Name: "availability-zone", Values: state.availabilityZoneNames },
        { Name: "default-for-az", Values: ["true"] },
      ],
    }),
  );
  state.subnets = Subnets.map((subnet) => subnet.SubnetId);
}),
new ScenarioOutput(
  "gotSubnets",
  /**
   * @param {{ subnets: string[] }} state
   */
  (state) =>
    MESSAGES.gotSubnets.replace("${SUBNETS}", state.subnets.join(", ")),
),
new ScenarioOutput(
  "creatingLoadBalancerTargetGroup",
  MESSAGES.creatingLoadBalancerTargetGroup.replace(
    "${TARGET_GROUP_NAME}",
    NAMES.loadBalancerTargetGroupName,
  ),
),
new ScenarioAction("createLoadBalancerTargetGroup", async (state) => {
  const client = new ElasticLoadBalancingV2Client({});
  const { TargetGroups } = await client.send(
    new CreateTargetGroupCommand({
      Name: NAMES.loadBalancerTargetGroupName,
      Protocol: "HTTP",
      Port: 80,
      HealthCheckPath: "/healthcheck",
      HealthCheckIntervalSeconds: 10,
      HealthCheckTimeoutSeconds: 5,
      HealthyThresholdCount: 2,
      UnhealthyThresholdCount: 2,
      VpcId: state.defaultVpc,
    }),
  );
  const targetGroup = TargetGroups[0];
  state.targetGroupArn = targetGroup.TargetGroupArn;
  state.targetGroupProtocol = targetGroup.Protocol;
  state.targetGroupPort = targetGroup.Port;
```

```
}),
new ScenarioOutput(
  "createdLoadBalancerTargetGroup",
  MESSAGES.createdLoadBalancerTargetGroup.replace(
    "${TARGET_GROUP_NAME}",
    NAMES.loadBalancerTargetGroupName,
  ),
),
new ScenarioOutput(
  "creatingLoadBalancer",
  MESSAGES.creatingLoadBalancer.replace("${LB_NAME}", NAMES.loadBalancerName),
),
new ScenarioAction("createLoadBalancer", async (state) => {
  const client = new ElasticLoadBalancingV2Client({});
  const { LoadBalancers } = await client.send(
    new CreateLoadBalancerCommand({
      Name: NAMES.loadBalancerName,
      Subnets: state.subnets,
    }),
  );
  state.loadBalancerDns = LoadBalancers[0].DNSName;
  state.loadBalancerArn = LoadBalancers[0].LoadBalancerArn;
  await waitUntilLoadBalancerAvailable(
    { client },
    { Names: [NAMES.loadBalancerName] },
  );
}),
new ScenarioOutput("createdLoadBalancer", (state) =>
  MESSAGES.createdLoadBalancer
    .replace("${LB_NAME}", NAMES.loadBalancerName)
    .replace("${DNS_NAME}", state.loadBalancerDns),
),
new ScenarioOutput(
  "creatingListener",
  MESSAGES.creatingLoadBalancerListener
    .replace("${LB_NAME}", NAMES.loadBalancerName)
    .replace("${TARGET_GROUP_NAME}", NAMES.loadBalancerTargetGroupName),
),
new ScenarioAction("createListener", async (state) => {
  const client = new ElasticLoadBalancingV2Client({});
  const { Listeners } = await client.send(
    new CreateListenerCommand({
      LoadBalancerArn: state.loadBalancerArn,
      Protocol: state.targetGroupProtocol,
```

```
        Port: state.targetGroupPort,
        DefaultActions: [
          { Type: "forward", TargetGroupArn: state.targetGroupArn },
        ],
      }),
    );
    const listener = Listeners[0];
    state.loadBalancerListenerArn = listener.ListenerArn;
  }),
  new ScenarioOutput("createdListener", (state) =>
    MESSAGES.createdLoadBalancerListener.replace(
      "${LB_LISTENER_ARN}",
      state.loadBalancerListenerArn,
    ),
  ),
  new ScenarioOutput(
    "attachingLoadBalancerTargetGroup",
    MESSAGES.attachingLoadBalancerTargetGroup
      .replace("${TARGET_GROUP_NAME}", NAMES.loadBalancerTargetGroupName)
      .replace("${AUTO_SCALING_GROUP_NAME}", NAMES.autoScalingGroupName),
  ),
  new ScenarioAction("attachLoadBalancerTargetGroup", async (state) => {
    const client = new AutoScalingClient({});
    await client.send(
      new AttachLoadBalancerTargetGroupsCommand({
        AutoScalingGroupName: NAMES.autoScalingGroupName,
        TargetGroupARNs: [state.targetGroupArn],
      }),
    );
  }),
  new ScenarioOutput(
    "attachedLoadBalancerTargetGroup",
    MESSAGES.attachedLoadBalancerTargetGroup,
  ),
  new ScenarioOutput("verifyingInboundPort", MESSAGES.verifyingInboundPort),
  new ScenarioAction(
    "verifyInboundPort",
    /**
     *
     * @param {{ defaultSecurityGroup: import('@aws-sdk/client-
ec2').SecurityGroup}} state
     */
    async (state) => {
      const client = new EC2Client({});
```

```
      const { SecurityGroups } = await client.send(
        new DescribeSecurityGroupsCommand({
          Filters: [{ Name: "group-name", Values: ["default"] }],
        }),
      );
      if (!SecurityGroups) {
        state.verifyInboundPortError = new Error(MESSAGES.noSecurityGroups);
      }
      state.defaultSecurityGroup = SecurityGroups[0];

      /**
       * @type {string}
       */
      const ipResponse = (await axios.get("http://checkip.amazonaws.com")).data;
      state.myIp = ipResponse.trim();
      const myIpRules = state.defaultSecurityGroup.IpPermissions.filter(
        ({ IpRanges }) =>
          IpRanges.some(
            ({ CidrIp }) =>
              CidrIp.startsWith(state.myIp) || CidrIp === "0.0.0.0/0",
          ),
      )
        .filter(({ IpProtocol }) => IpProtocol === "tcp")
        .filter(({ FromPort }) => FromPort === 80);

      state.myIpRules = myIpRules;
    },
  ),
  new ScenarioOutput(
    "verifiedInboundPort",
    /**
     * @param {{ myIpRules: any[] }} state
     */
    (state) => {
      if (state.myIpRules.length > 0) {
        return MESSAGES.foundIpRules.replace(
          "${IP_RULES}",
          JSON.stringify(state.myIpRules, null, 2),
        );
      }
      return MESSAGES.noIpRules;
    },
  ),
  new ScenarioInput(
```

```
    "shouldAddInboundRule",
    /**
     * @param {{ myIpRules: any[] }} state
     */
    (state) => {
      if (state.myIpRules.length > 0) {
        return false;
      }
      return MESSAGES.noIpRules;
    },
    { type: "confirm" },
  ),
  new ScenarioAction(
    "addInboundRule",
    /**
     * @param {{ defaultSecurityGroup: import('@aws-sdk/client-
ec2').SecurityGroup }} state
     */
    async (state) => {
      if (!state.shouldAddInboundRule) {
        return;
      }

      const client = new EC2Client({});
      await client.send(
        new AuthorizeSecurityGroupIngressCommand({
          GroupId: state.defaultSecurityGroup.GroupId,
          CidrIp: `${state.myIp}/32`,
          FromPort: 80,
          ToPort: 80,
          IpProtocol: "tcp",
        }),
      );
    },
  ),
  new ScenarioOutput("addedInboundRule", (state) => {
    if (state.shouldAddInboundRule) {
      return MESSAGES.addedInboundRule.replace("${IP_ADDRESS}", state.myIp);
    }
    return false;
  }),
  new ScenarioOutput("verifyingEndpoint", (state) =>
    MESSAGES.verifyingEndpoint.replace("${DNS_NAME}", state.loadBalancerDns),
  ),
```

```
  new ScenarioAction("verifyEndpoint", async (state) => {
    try {
      const response = await retry({ intervalInMs: 2000, maxRetries: 30 }, () =>
        axios.get(`http://${state.loadBalancerDns}`),
      );
      state.endpointResponse = JSON.stringify(response.data, null, 2);
    } catch (e) {
      state.verifyEndpointError = e;
    }
  }),
  new ScenarioOutput("verifiedEndpoint", (state) => {
    if (state.verifyEndpointError) {
      console.error(state.verifyEndpointError);
    } else {
      return MESSAGES.verifiedEndpoint.replace(
        "${ENDPOINT_RESPONSE}",
        state.endpointResponse,
      );
    }
  }),
  saveState,
];
```

Criar etapas para executar a demonstração.

```
import { readFileSync } from "node:fs";
import { join } from "node:path";

import axios from "axios";

import {
  DescribeTargetGroupsCommand,
  DescribeTargetHealthCommand,
  ElasticLoadBalancingV2Client,
} from "@aws-sdk/client-elastic-load-balancing-v2";
import {
  DescribeInstanceInformationCommand,
  PutParameterCommand,
  SSMClient,
  SendCommandCommand,
} from "@aws-sdk/client-ssm";
import {
```

```
  IAMClient,
  CreatePolicyCommand,
  CreateRoleCommand,
  AttachRolePolicyCommand,
  CreateInstanceProfileCommand,
  AddRoleToInstanceProfileCommand,
  waitUntilInstanceProfileExists,
} from "@aws-sdk/client-iam";
import {
  AutoScalingClient,
  DescribeAutoScalingGroupsCommand,
  TerminateInstanceInAutoScalingGroupCommand,
} from "@aws-sdk/client-auto-scaling";
import {
  DescribeIamInstanceProfileAssociationsCommand,
  EC2Client,
  RebootInstancesCommand,
  ReplaceIamInstanceProfileAssociationCommand,
} from "@aws-sdk/client-ec2";

import {
  ScenarioAction,
  ScenarioInput,
  ScenarioOutput,
} from "@aws-doc-sdk-examples/lib/scenario/scenario.js";
import { retry } from "@aws-doc-sdk-examples/lib/utils/util-timers.js";

import { MESSAGES, NAMES, RESOURCES_PATH } from "./constants.js";
import { findLoadBalancer } from "./shared.js";

const getRecommendation = new ScenarioAction(
  "getRecommendation",
  async (state) => {
    const loadBalancer = await findLoadBalancer(NAMES.loadBalancerName);
    if (loadBalancer) {
      state.loadBalancerDnsName = loadBalancer.DNSName;
      try {
        state.recommendation = (
          await axios.get(`http://${state.loadBalancerDnsName}`)
        ).data;
      } catch (e) {
        state.recommendation = e instanceof Error ? e.message : e;
      }
    } else {
```

```
      throw new Error(MESSAGES.demoFindLoadBalancerError);
    }
  },
);

const getRecommendationResult = new ScenarioOutput(
  "getRecommendationResult",
  (state) =>
    `Recommendation:\n${JSON.stringify(state.recommendation, null, 2)}`,
  { preformatted: true },
);

const getHealthCheck = new ScenarioAction("getHealthCheck", async (state) => {
  const client = new ElasticLoadBalancingV2Client({});
  const { TargetGroups } = await client.send(
    new DescribeTargetGroupsCommand({
      Names: [NAMES.loadBalancerTargetGroupName],
    }),
  );

  const { TargetHealthDescriptions } = await client.send(
    new DescribeTargetHealthCommand({
      TargetGroupArn: TargetGroups[0].TargetGroupArn,
    }),
  );
  state.targetHealthDescriptions = TargetHealthDescriptions;
});

const getHealthCheckResult = new ScenarioOutput(
  "getHealthCheckResult",
  /**
   * @param {{ targetHealthDescriptions: import('@aws-sdk/client-elastic-load-
balancing-v2').TargetHealthDescription[]}} state
   */
  (state) => {
    const status = state.targetHealthDescriptions
      .map((th) => `${th.Target.Id}: ${th.TargetHealth.State}`)
      .join("\n");
    return `Health check:\n${status}`;
  },
  { preformatted: true },
);

const loadBalancerLoop = new ScenarioAction(
```

```
  "loadBalancerLoop",
  getRecommendation.action,
  {
    whileConfig: {
      whileFn: ({ loadBalancerCheck }) => loadBalancerCheck,
      input: new ScenarioInput(
        "loadBalancerCheck",
        MESSAGES.demoLoadBalancerCheck,
        {
          type: "confirm",
        },
      ),
      output: getRecommendationResult,
    },
  },
);

const healthCheckLoop = new ScenarioAction(
  "healthCheckLoop",
  getHealthCheck.action,
  {
    whileConfig: {
      whileFn: ({ healthCheck }) => healthCheck,
      input: new ScenarioInput("healthCheck", MESSAGES.demoHealthCheck, {
        type: "confirm",
      }),
      output: getHealthCheckResult,
    },
  },
);

const statusSteps = [
  getRecommendation,
  getRecommendationResult,
  getHealthCheck,
  getHealthCheckResult,
];

/**
 * @type {import('@aws-doc-sdk-examples/lib/scenario.js').Step[]}
 */
export const demoSteps = [
  new ScenarioOutput("header", MESSAGES.demoHeader, { header: true }),
  new ScenarioOutput("sanityCheck", MESSAGES.demoSanityCheck),
```

```
...statusSteps,
new ScenarioInput(
  "brokenDependencyConfirmation",
  MESSAGES.demoBrokenDependencyConfirmation,
  { type: "confirm" },
),
new ScenarioAction("brokenDependency", async (state) => {
  if (!state.brokenDependencyConfirmation) {
    process.exit();
  } else {
    const client = new SSMClient({});
    state.badTableName = `fake-table-${Date.now()}`;
    await client.send(
      new PutParameterCommand({
        Name: NAMES.ssmTableNameKey,
        Value: state.badTableName,
        Overwrite: true,
        Type: "String",
      }),
    );
  }
}),
new ScenarioOutput("testBrokenDependency", (state) =>
  MESSAGES.demoTestBrokenDependency.replace(
    "${TABLE_NAME}",
    state.badTableName,
  ),
),
...statusSteps,
new ScenarioInput(
  "staticResponseConfirmation",
  MESSAGES.demoStaticResponseConfirmation,
  { type: "confirm" },
),
new ScenarioAction("staticResponse", async (state) => {
  if (!state.staticResponseConfirmation) {
    process.exit();
  } else {
    const client = new SSMClient({});
    await client.send(
      new PutParameterCommand({
        Name: NAMES.ssmFailureResponseKey,
        Value: "static",
        Overwrite: true,
```

```
          Type: "String",
        }),
      );
    }
  }),
  new ScenarioOutput("testStaticResponse", MESSAGES.demoTestStaticResponse),
  ...statusSteps,
  new ScenarioInput(
    "badCredentialsConfirmation",
    MESSAGES.demoBadCredentialsConfirmation,
    { type: "confirm" },
  ),
  new ScenarioAction("badCredentialsExit", (state) => {
    if (!state.badCredentialsConfirmation) {
      process.exit();
    }
  }),
  new ScenarioAction("fixDynamoDBName", async () => {
    const client = new SSMClient({});
    await client.send(
      new PutParameterCommand({
        Name: NAMES.ssmTableNameKey,
        Value: NAMES.tableName,
        Overwrite: true,
        Type: "String",
      }),
    );
  }),
  new ScenarioAction(
    "badCredentials",
    /**
     * @param {{ targetInstance: import('@aws-sdk/client-auto-
scaling').Instance }} state
     */
    async (state) => {
      await createSsmOnlyInstanceProfile();
      const autoScalingClient = new AutoScalingClient({});
      const { AutoScalingGroups } = await autoScalingClient.send(
        new DescribeAutoScalingGroupsCommand({
          AutoScalingGroupNames: [NAMES.autoScalingGroupName],
        }),
      );
      state.targetInstance = AutoScalingGroups[0].Instances[0];
      const ec2Client = new EC2Client({});
```

```
      const { IamInstanceProfileAssociations } = await ec2Client.send(
        new DescribeIamInstanceProfileAssociationsCommand({
          Filters: [
            { Name: "instance-id", Values: [state.targetInstance.InstanceId] },
          ],
        }),
      );
      state.instanceProfileAssociationId =
        IamInstanceProfileAssociations[0].AssociationId;
      await retry({ intervalInMs: 1000, maxRetries: 30 }, () =>
        ec2Client.send(
          new ReplaceIamInstanceProfileAssociationCommand({
            AssociationId: state.instanceProfileAssociationId,
            IamInstanceProfile: { Name: NAMES.ssmOnlyInstanceProfileName },
          }),
        ),
      );

      await ec2Client.send(
        new RebootInstancesCommand({
          InstanceIds: [state.targetInstance.InstanceId],
        }),
      );

      const ssmClient = new SSMClient({});
      await retry({ intervalInMs: 20000, maxRetries: 15 }, async () => {
        const { InstanceInformationList } = await ssmClient.send(
          new DescribeInstanceInformationCommand({}),
        );

        const instance = InstanceInformationList.find(
          (info) => info.InstanceId === state.targetInstance.InstanceId,
        );

        if (!instance) {
          throw new Error("Instance not found.");
        }
      });

      await ssmClient.send(
        new SendCommandCommand({
          InstanceIds: [state.targetInstance.InstanceId],
          DocumentName: "AWS-RunShellScript",
          Parameters: { commands: ["cd / && sudo python3 server.py 80"] },
```

```
        }),
      );
    },
  ),
  new ScenarioOutput(
    "testBadCredentials",
    /**
     * @param {{ targetInstance: import('@aws-sdk/client-
ssm').InstanceInformation}} state
     */
    (state) =>
      MESSAGES.demoTestBadCredentials.replace(
        "${INSTANCE_ID}",
        state.targetInstance.InstanceId,
      ),
  ),
  loadBalancerLoop,
  new ScenarioInput(
    "deepHealthCheckConfirmation",
    MESSAGES.demoDeepHealthCheckConfirmation,
    { type: "confirm" },
  ),
  new ScenarioAction("deepHealthCheckExit", (state) => {
    if (!state.deepHealthCheckConfirmation) {
      process.exit();
    }
  }),
  new ScenarioAction("deepHealthCheck", async () => {
    const client = new SSMClient({});
    await client.send(
      new PutParameterCommand({
        Name: NAMES.ssmHealthCheckKey,
        Value: "deep",
        Overwrite: true,
        Type: "String",
      }),
    );
  }),
  new ScenarioOutput("testDeepHealthCheck", MESSAGES.demoTestDeepHealthCheck),
  healthCheckLoop,
  loadBalancerLoop,
  new ScenarioInput(
    "killInstanceConfirmation",
    /**
```

```
     * @param {{ targetInstance: import('@aws-sdk/client-
ssm').InstanceInformation }} state
     */
    (state) =>
      MESSAGES.demoKillInstanceConfirmation.replace(
        "${INSTANCE_ID}",
        state.targetInstance.InstanceId,
      ),
    { type: "confirm" },
  ),
  new ScenarioAction("killInstanceExit", (state) => {
    if (!state.killInstanceConfirmation) {
      process.exit();
    }
  }),
  new ScenarioAction(
    "killInstance",
    /**
     * @param {{ targetInstance: import('@aws-sdk/client-
ssm').InstanceInformation }} state
     */
    async (state) => {
      const client = new AutoScalingClient({});
      await client.send(
        new TerminateInstanceInAutoScalingGroupCommand({
          InstanceId: state.targetInstance.InstanceId,
          ShouldDecrementDesiredCapacity: false,
        }),
      );
    },
  ),
  new ScenarioOutput("testKillInstance", MESSAGES.demoTestKillInstance),
  healthCheckLoop,
  loadBalancerLoop,
  new ScenarioInput("failOpenConfirmation", MESSAGES.demoFailOpenConfirmation, {
    type: "confirm",
  }),
  new ScenarioAction("failOpenExit", (state) => {
    if (!state.failOpenConfirmation) {
      process.exit();
    }
  }),
  new ScenarioAction("failOpen", () => {
    const client = new SSMClient({});
```

```
    return client.send(
      new PutParameterCommand({
        Name: NAMES.ssmTableNameKey,
        Value: `fake-table-${Date.now()}`,
        Overwrite: true,
        Type: "String",
      }),
    );
  }),
  new ScenarioOutput("testFailOpen", MESSAGES.demoFailOpenTest),
  healthCheckLoop,
  loadBalancerLoop,
  new ScenarioInput(
    "resetTableConfirmation",
    MESSAGES.demoResetTableConfirmation,
    { type: "confirm" },
  ),
  new ScenarioAction("resetTableExit", (state) => {
    if (!state.resetTableConfirmation) {
      process.exit();
    }
  }),
  new ScenarioAction("resetTable", async () => {
    const client = new SSMClient({});
    await client.send(
      new PutParameterCommand({
        Name: NAMES.ssmTableNameKey,
        Value: NAMES.tableName,
        Overwrite: true,
        Type: "String",
      }),
    );
  }),
  new ScenarioOutput("testResetTable", MESSAGES.demoTestResetTable),
  healthCheckLoop,
  loadBalancerLoop,
];

async function createSsmOnlyInstanceProfile() {
  const iamClient = new IAMClient({});
  const { Policy } = await iamClient.send(
    new CreatePolicyCommand({
      PolicyName: NAMES.ssmOnlyPolicyName,
      PolicyDocument: readFileSync(
```

```
      join(RESOURCES_PATH, "ssm_only_policy.json"),
    ),
  }),
);
await iamClient.send(
  new CreateRoleCommand({
    RoleName: NAMES.ssmOnlyRoleName,
    AssumeRolePolicyDocument: JSON.stringify({
      Version: "2012-10-17",
      Statement: [
        {
          Effect: "Allow",
          Principal: { Service: "ec2.amazonaws.com" },
          Action: "sts:AssumeRole",
        },
      ],
    }),
  }),
);
await iamClient.send(
  new AttachRolePolicyCommand({
    RoleName: NAMES.ssmOnlyRoleName,
    PolicyArn: Policy.Arn,
  }),
);
await iamClient.send(
  new AttachRolePolicyCommand({
    RoleName: NAMES.ssmOnlyRoleName,
    PolicyArn: "arn:aws:iam::aws:policy/AmazonSSMManagedInstanceCore",
  }),
);
const { InstanceProfile } = await iamClient.send(
  new CreateInstanceProfileCommand({
    InstanceProfileName: NAMES.ssmOnlyInstanceProfileName,
  }),
);
await waitUntilInstanceProfileExists(
  { client: iamClient },
  { InstanceProfileName: NAMES.ssmOnlyInstanceProfileName },
);
await iamClient.send(
  new AddRoleToInstanceProfileCommand({
    InstanceProfileName: NAMES.ssmOnlyInstanceProfileName,
    RoleName: NAMES.ssmOnlyRoleName,
```

```
    }),
  );

  return InstanceProfile;
}
```

Criar etapas para destruir todos os recursos.

```
import { unlinkSync } from "node:fs";

import { DynamoDBClient, DeleteTableCommand } from "@aws-sdk/client-dynamodb";
import {
  EC2Client,
  DeleteKeyPairCommand,
  DeleteLaunchTemplateCommand,
  RevokeSecurityGroupIngressCommand,
} from "@aws-sdk/client-ec2";
import {
  IAMClient,
  DeleteInstanceProfileCommand,
  RemoveRoleFromInstanceProfileCommand,
  DeletePolicyCommand,
  DeleteRoleCommand,
  DetachRolePolicyCommand,
  paginateListPolicies,
} from "@aws-sdk/client-iam";
import {
  AutoScalingClient,
  DeleteAutoScalingGroupCommand,
  TerminateInstanceInAutoScalingGroupCommand,
  UpdateAutoScalingGroupCommand,
  paginateDescribeAutoScalingGroups,
} from "@aws-sdk/client-auto-scaling";
import {
  DeleteLoadBalancerCommand,
  DeleteTargetGroupCommand,
  DescribeTargetGroupsCommand,
  ElasticLoadBalancingV2Client,
} from "@aws-sdk/client-elastic-load-balancing-v2";

import {
  ScenarioOutput,
```

```
  ScenarioInput,
  ScenarioAction,
} from "@aws-doc-sdk-examples/lib/scenario/index.js";
import { loadState } from "@aws-doc-sdk-examples/lib/scenario/steps-common.js";
import { retry } from "@aws-doc-sdk-examples/lib/utils/util-timers.js";

import { MESSAGES, NAMES } from "./constants.js";
import { findLoadBalancer } from "./shared.js";

/**
 * @type {import('@aws-doc-sdk-examples/lib/scenario.js').Step[]}
 */
export const destroySteps = [
  loadState,
  new ScenarioInput("destroy", MESSAGES.destroy, { type: "confirm" }),
  new ScenarioAction(
    "abort",
    (state) => state.destroy === false && process.exit(),
  ),
  new ScenarioAction("deleteTable", async (c) => {
    try {
      const client = new DynamoDBClient({});
      await client.send(new DeleteTableCommand({ TableName: NAMES.tableName }));
    } catch (e) {
      c.deleteTableError = e;
    }
  }),
  new ScenarioOutput("deleteTableResult", (state) => {
    if (state.deleteTableError) {
      console.error(state.deleteTableError);
      return MESSAGES.deleteTableError.replace(
        "${TABLE_NAME}",
        NAMES.tableName,
      );
    }
    return MESSAGES.deletedTable.replace("${TABLE_NAME}", NAMES.tableName);
  }),
  new ScenarioAction("deleteKeyPair", async (state) => {
    try {
      const client = new EC2Client({});
      await client.send(
        new DeleteKeyPairCommand({ KeyName: NAMES.keyPairName }),
      );
      unlinkSync(`${NAMES.keyPairName}.pem`);
```

```
  } catch (e) {
    state.deleteKeyPairError = e;
  }
}),
new ScenarioOutput("deleteKeyPairResult", (state) => {
  if (state.deleteKeyPairError) {
    console.error(state.deleteKeyPairError);
    return MESSAGES.deleteKeyPairError.replace(
      "${KEY_PAIR_NAME}",
      NAMES.keyPairName,
    );
  }
  return MESSAGES.deletedKeyPair.replace(
    "${KEY_PAIR_NAME}",
    NAMES.keyPairName,
  );
}),
new ScenarioAction("detachPolicyFromRole", async (state) => {
  try {
    const client = new IAMClient({});
    const policy = await findPolicy(NAMES.instancePolicyName);

    if (!policy) {
      state.detachPolicyFromRoleError = new Error(
        `Policy ${NAMES.instancePolicyName} not found.`,
      );
    } else {
      await client.send(
        new DetachRolePolicyCommand({
          RoleName: NAMES.instanceRoleName,
          PolicyArn: policy.Arn,
        }),
      );
    }
  } catch (e) {
    state.detachPolicyFromRoleError = e;
  }
}),
new ScenarioOutput("detachedPolicyFromRole", (state) => {
  if (state.detachPolicyFromRoleError) {
    console.error(state.detachPolicyFromRoleError);
    return MESSAGES.detachPolicyFromRoleError
      .replace("${INSTANCE_POLICY_NAME}", NAMES.instancePolicyName)
      .replace("${INSTANCE_ROLE_NAME}", NAMES.instanceRoleName);
```

```
    }
    return MESSAGES.detachedPolicyFromRole
      .replace("${INSTANCE_POLICY_NAME}", NAMES.instancePolicyName)
      .replace("${INSTANCE_ROLE_NAME}", NAMES.instanceRoleName);
  }),
  new ScenarioAction("deleteInstancePolicy", async (state) => {
    const client = new IAMClient({});
    const policy = await findPolicy(NAMES.instancePolicyName);

    if (!policy) {
      state.deletePolicyError = new Error(
        `Policy ${NAMES.instancePolicyName} not found.`,
      );
    } else {
      return client.send(
        new DeletePolicyCommand({
          PolicyArn: policy.Arn,
        }),
      );
    }
  }),
  new ScenarioOutput("deletePolicyResult", (state) => {
    if (state.deletePolicyError) {
      console.error(state.deletePolicyError);
      return MESSAGES.deletePolicyError.replace(
        "${INSTANCE_POLICY_NAME}",
        NAMES.instancePolicyName,
      );
    }
    return MESSAGES.deletedPolicy.replace(
      "${INSTANCE_POLICY_NAME}",
      NAMES.instancePolicyName,
    );
  }),
  new ScenarioAction("removeRoleFromInstanceProfile", async (state) => {
    try {
      const client = new IAMClient({});
      await client.send(
        new RemoveRoleFromInstanceProfileCommand({
          RoleName: NAMES.instanceRoleName,
          InstanceProfileName: NAMES.instanceProfileName,
        }),
      );
    } catch (e) {
```

```
      state.removeRoleFromInstanceProfileError = e;
    }
  }),
  new ScenarioOutput("removeRoleFromInstanceProfileResult", (state) => {
    if (state.removeRoleFromInstanceProfile) {
      console.error(state.removeRoleFromInstanceProfileError);
      return MESSAGES.removeRoleFromInstanceProfileError
        .replace("${INSTANCE_PROFILE_NAME}", NAMES.instanceProfileName)
        .replace("${INSTANCE_ROLE_NAME}", NAMES.instanceRoleName);
    }
    return MESSAGES.removedRoleFromInstanceProfile
      .replace("${INSTANCE_PROFILE_NAME}", NAMES.instanceProfileName)
      .replace("${INSTANCE_ROLE_NAME}", NAMES.instanceRoleName);
  }),
  new ScenarioAction("deleteInstanceRole", async (state) => {
    try {
      const client = new IAMClient({});
      await client.send(
        new DeleteRoleCommand({
          RoleName: NAMES.instanceRoleName,
        }),
      );
    } catch (e) {
      state.deleteInstanceRoleError = e;
    }
  }),
  new ScenarioOutput("deleteInstanceRoleResult", (state) => {
    if (state.deleteInstanceRoleError) {
      console.error(state.deleteInstanceRoleError);
      return MESSAGES.deleteInstanceRoleError.replace(
        "${INSTANCE_ROLE_NAME}",
        NAMES.instanceRoleName,
      );
    }
    return MESSAGES.deletedInstanceRole.replace(
      "${INSTANCE_ROLE_NAME}",
      NAMES.instanceRoleName,
    );
  }),
  new ScenarioAction("deleteInstanceProfile", async (state) => {
    try {
      const client = new IAMClient({});
      await client.send(
        new DeleteInstanceProfileCommand({
```

```
          InstanceProfileName: NAMES.instanceProfileName,
        }),
      );
    } catch (e) {
      state.deleteInstanceProfileError = e;
    }
  }),
  new ScenarioOutput("deleteInstanceProfileResult", (state) => {
    if (state.deleteInstanceProfileError) {
      console.error(state.deleteInstanceProfileError);
      return MESSAGES.deleteInstanceProfileError.replace(
        "${INSTANCE_PROFILE_NAME}",
        NAMES.instanceProfileName,
      );
    }
    return MESSAGES.deletedInstanceProfile.replace(
      "${INSTANCE_PROFILE_NAME}",
      NAMES.instanceProfileName,
    );
  }),
  new ScenarioAction("deleteAutoScalingGroup", async (state) => {
    try {
      await terminateGroupInstances(NAMES.autoScalingGroupName);
      await retry({ intervalInMs: 60000, maxRetries: 60 }, async () => {
        await deleteAutoScalingGroup(NAMES.autoScalingGroupName);
      });
    } catch (e) {
      state.deleteAutoScalingGroupError = e;
    }
  }),
  new ScenarioOutput("deleteAutoScalingGroupResult", (state) => {
    if (state.deleteAutoScalingGroupError) {
      console.error(state.deleteAutoScalingGroupError);
      return MESSAGES.deleteAutoScalingGroupError.replace(
        "${AUTO_SCALING_GROUP_NAME}",
        NAMES.autoScalingGroupName,
      );
    }
    return MESSAGES.deletedAutoScalingGroup.replace(
      "${AUTO_SCALING_GROUP_NAME}",
      NAMES.autoScalingGroupName,
    );
  }),
  new ScenarioAction("deleteLaunchTemplate", async (state) => {
```

```
    const client = new EC2Client({});
    try {
      await client.send(
        new DeleteLaunchTemplateCommand({
          LaunchTemplateName: NAMES.launchTemplateName,
        }),
      );
    } catch (e) {
      state.deleteLaunchTemplateError = e;
    }
  }),
  new ScenarioOutput("deleteLaunchTemplateResult", (state) => {
    if (state.deleteLaunchTemplateError) {
      console.error(state.deleteLaunchTemplateError);
      return MESSAGES.deleteLaunchTemplateError.replace(
        "${LAUNCH_TEMPLATE_NAME}",
        NAMES.launchTemplateName,
      );
    }
    return MESSAGES.deletedLaunchTemplate.replace(
      "${LAUNCH_TEMPLATE_NAME}",
      NAMES.launchTemplateName,
    );
  }),
  new ScenarioAction("deleteLoadBalancer", async (state) => {
    try {
      const client = new ElasticLoadBalancingV2Client({});
      const loadBalancer = await findLoadBalancer(NAMES.loadBalancerName);
      await client.send(
        new DeleteLoadBalancerCommand({
          LoadBalancerArn: loadBalancer.LoadBalancerArn,
        }),
      );
      await retry({ intervalInMs: 1000, maxRetries: 60 }, async () => {
        const lb = await findLoadBalancer(NAMES.loadBalancerName);
        if (lb) {
          throw new Error("Load balancer still exists.");
        }
      });
    } catch (e) {
      state.deleteLoadBalancerError = e;
    }
  }),
  new ScenarioOutput("deleteLoadBalancerResult", (state) => {
```

```
  if (state.deleteLoadBalancerError) {
    console.error(state.deleteLoadBalancerError);
    return MESSAGES.deleteLoadBalancerError.replace(
      "${LB_NAME}",
      NAMES.loadBalancerName,
    );
  }
  return MESSAGES.deletedLoadBalancer.replace(
    "${LB_NAME}",
    NAMES.loadBalancerName,
  );
}),
new ScenarioAction("deleteLoadBalancerTargetGroup", async (state) => {
  const client = new ElasticLoadBalancingV2Client({});
  try {
    const { TargetGroups } = await client.send(
      new DescribeTargetGroupsCommand({
        Names: [NAMES.loadBalancerTargetGroupName],
      }),
    );

    await retry({ intervalInMs: 1000, maxRetries: 30 }, () =>
      client.send(
        new DeleteTargetGroupCommand({
          TargetGroupArn: TargetGroups[0].TargetGroupArn,
        }),
      ),
    );
  } catch (e) {
    state.deleteLoadBalancerTargetGroupError = e;
  }
}),
new ScenarioOutput("deleteLoadBalancerTargetGroupResult", (state) => {
  if (state.deleteLoadBalancerTargetGroupError) {
    console.error(state.deleteLoadBalancerTargetGroupError);
    return MESSAGES.deleteLoadBalancerTargetGroupError.replace(
      "${TARGET_GROUP_NAME}",
      NAMES.loadBalancerTargetGroupName,
    );
  }
  return MESSAGES.deletedLoadBalancerTargetGroup.replace(
    "${TARGET_GROUP_NAME}",
    NAMES.loadBalancerTargetGroupName,
  );
```

```
  }),
  new ScenarioAction("detachSsmOnlyRoleFromProfile", async (state) => {
    try {
      const client = new IAMClient({});
      await client.send(
        new RemoveRoleFromInstanceProfileCommand({
          InstanceProfileName: NAMES.ssmOnlyInstanceProfileName,
          RoleName: NAMES.ssmOnlyRoleName,
        }),
      );
    } catch (e) {
      state.detachSsmOnlyRoleFromProfileError = e;
    }
  }),
  new ScenarioOutput("detachSsmOnlyRoleFromProfileResult", (state) => {
    if (state.detachSsmOnlyRoleFromProfileError) {
      console.error(state.detachSsmOnlyRoleFromProfileError);
      return MESSAGES.detachSsmOnlyRoleFromProfileError
        .replace("${ROLE_NAME}", NAMES.ssmOnlyRoleName)
        .replace("${PROFILE_NAME}", NAMES.ssmOnlyInstanceProfileName);
    }
    return MESSAGES.detachedSsmOnlyRoleFromProfile
      .replace("${ROLE_NAME}", NAMES.ssmOnlyRoleName)
      .replace("${PROFILE_NAME}", NAMES.ssmOnlyInstanceProfileName);
  }),
  new ScenarioAction("detachSsmOnlyCustomRolePolicy", async (state) => {
    try {
      const iamClient = new IAMClient({});
      const ssmOnlyPolicy = await findPolicy(NAMES.ssmOnlyPolicyName);
      await iamClient.send(
        new DetachRolePolicyCommand({
          RoleName: NAMES.ssmOnlyRoleName,
          PolicyArn: ssmOnlyPolicy.Arn,
        }),
      );
    } catch (e) {
      state.detachSsmOnlyCustomRolePolicyError = e;
    }
  }),
  new ScenarioOutput("detachSsmOnlyCustomRolePolicyResult", (state) => {
    if (state.detachSsmOnlyCustomRolePolicyError) {
      console.error(state.detachSsmOnlyCustomRolePolicyError);
      return MESSAGES.detachSsmOnlyCustomRolePolicyError
        .replace("${ROLE_NAME}", NAMES.ssmOnlyRoleName)
```

```
        .replace("${POLICY_NAME}", NAMES.ssmOnlyPolicyName);
    }
    return MESSAGES.detachedSsmOnlyCustomRolePolicy
      .replace("${ROLE_NAME}", NAMES.ssmOnlyRoleName)
      .replace("${POLICY_NAME}", NAMES.ssmOnlyPolicyName);
  }),
  new ScenarioAction("detachSsmOnlyAWSRolePolicy", async (state) => {
    try {
      const iamClient = new IAMClient({});
      await iamClient.send(
        new DetachRolePolicyCommand({
          RoleName: NAMES.ssmOnlyRoleName,
          PolicyArn: "arn:aws:iam::aws:policy/AmazonSSMManagedInstanceCore",
        }),
      );
    } catch (e) {
      state.detachSsmOnlyAWSRolePolicyError = e;
    }
  }),
  new ScenarioOutput("detachSsmOnlyAWSRolePolicyResult", (state) => {
    if (state.detachSsmOnlyAWSRolePolicyError) {
      console.error(state.detachSsmOnlyAWSRolePolicyError);
      return MESSAGES.detachSsmOnlyAWSRolePolicyError
        .replace("${ROLE_NAME}", NAMES.ssmOnlyRoleName)
        .replace("${POLICY_NAME}", "AmazonSSMManagedInstanceCore");
    }
    return MESSAGES.detachedSsmOnlyAWSRolePolicy
      .replace("${ROLE_NAME}", NAMES.ssmOnlyRoleName)
      .replace("${POLICY_NAME}", "AmazonSSMManagedInstanceCore");
  }),
  new ScenarioAction("deleteSsmOnlyInstanceProfile", async (state) => {
    try {
      const iamClient = new IAMClient({});
      await iamClient.send(
        new DeleteInstanceProfileCommand({
          InstanceProfileName: NAMES.ssmOnlyInstanceProfileName,
        }),
      );
    } catch (e) {
      state.deleteSsmOnlyInstanceProfileError = e;
    }
  }),
  new ScenarioOutput("deleteSsmOnlyInstanceProfileResult", (state) => {
    if (state.deleteSsmOnlyInstanceProfileError) {
```

```
      console.error(state.deleteSsmOnlyInstanceProfileError);
      return MESSAGES.deleteSsmOnlyInstanceProfileError.replace(
        "${INSTANCE_PROFILE_NAME}",
        NAMES.ssmOnlyInstanceProfileName,
      );
    }
    return MESSAGES.deletedSsmOnlyInstanceProfile.replace(
      "${INSTANCE_PROFILE_NAME}",
      NAMES.ssmOnlyInstanceProfileName,
    );
  }),
  new ScenarioAction("deleteSsmOnlyPolicy", async (state) => {
    try {
      const iamClient = new IAMClient({});
      const ssmOnlyPolicy = await findPolicy(NAMES.ssmOnlyPolicyName);
      await iamClient.send(
        new DeletePolicyCommand({
          PolicyArn: ssmOnlyPolicy.Arn,
        }),
      );
    } catch (e) {
      state.deleteSsmOnlyPolicyError = e;
    }
  }),
  new ScenarioOutput("deleteSsmOnlyPolicyResult", (state) => {
    if (state.deleteSsmOnlyPolicyError) {
      console.error(state.deleteSsmOnlyPolicyError);
      return MESSAGES.deleteSsmOnlyPolicyError.replace(
        "${POLICY_NAME}",
        NAMES.ssmOnlyPolicyName,
      );
    }
    return MESSAGES.deletedSsmOnlyPolicy.replace(
      "${POLICY_NAME}",
      NAMES.ssmOnlyPolicyName,
    );
  }),
  new ScenarioAction("deleteSsmOnlyRole", async (state) => {
    try {
      const iamClient = new IAMClient({});
      await iamClient.send(
        new DeleteRoleCommand({
          RoleName: NAMES.ssmOnlyRoleName,
        }),
```

```
      );
    } catch (e) {
      state.deleteSsmOnlyRoleError = e;
    }
  }),
  new ScenarioOutput("deleteSsmOnlyRoleResult", (state) => {
    if (state.deleteSsmOnlyRoleError) {
      console.error(state.deleteSsmOnlyRoleError);
      return MESSAGES.deleteSsmOnlyRoleError.replace(
        "${ROLE_NAME}",
        NAMES.ssmOnlyRoleName,
      );
    }
    return MESSAGES.deletedSsmOnlyRole.replace(
      "${ROLE_NAME}",
      NAMES.ssmOnlyRoleName,
    );
  }),
  new ScenarioAction(
    "revokeSecurityGroupIngress",
    async (
      /** @type {{ myIp: string, defaultSecurityGroup: { GroupId: string } }} */
state,
    ) => {
      const ec2Client = new EC2Client({});

      try {
        await ec2Client.send(
          new RevokeSecurityGroupIngressCommand({
            GroupId: state.defaultSecurityGroup.GroupId,
            CidrIp: `${state.myIp}/32`,
            FromPort: 80,
            ToPort: 80,
            IpProtocol: "tcp",
          }),
        );
      } catch (e) {
        state.revokeSecurityGroupIngressError = e;
      }
    },
  ),
  new ScenarioOutput("revokeSecurityGroupIngressResult", (state) => {
    if (state.revokeSecurityGroupIngressError) {
      console.error(state.revokeSecurityGroupIngressError);
```

```
      return MESSAGES.revokeSecurityGroupIngressError.replace(
        "${IP}",
        state.myIp,
      );
    }
    return MESSAGES.revokedSecurityGroupIngress.replace("${IP}", state.myIp);
  }),
];

/**
 * @param {string} policyName
 */
async function findPolicy(policyName) {
  const client = new IAMClient({});
  const paginatedPolicies = paginateListPolicies({ client }, {});
  for await (const page of paginatedPolicies) {
    const policy = page.Policies.find((p) => p.PolicyName === policyName);
    if (policy) {
      return policy;
    }
  }
}

/**
 * @param {string} groupName
 */
async function deleteAutoScalingGroup(groupName) {
  const client = new AutoScalingClient({});
  try {
    await client.send(
      new DeleteAutoScalingGroupCommand({
        AutoScalingGroupName: groupName,
      }),
    );
  } catch (err) {
    if (!(err instanceof Error)) {
      throw err;
    }
    console.log(err.name);
    throw err;
  }
}

/**
```

```
 * @param {string} groupName
 */
async function terminateGroupInstances(groupName) {
  const autoScalingClient = new AutoScalingClient({});
  const group = await findAutoScalingGroup(groupName);
  await autoScalingClient.send(
    new UpdateAutoScalingGroupCommand({
      AutoScalingGroupName: group.AutoScalingGroupName,
      MinSize: 0,
    }),
  );
  for (const i of group.Instances) {
    await retry({ intervalInMs: 1000, maxRetries: 30 }, () =>
      autoScalingClient.send(
        new TerminateInstanceInAutoScalingGroupCommand({
          InstanceId: i.InstanceId,
          ShouldDecrementDesiredCapacity: true,
        }),
      ),
    );
  }
}

async function findAutoScalingGroup(groupName) {
  const client = new AutoScalingClient({});
  const paginatedGroups = paginateDescribeAutoScalingGroups({ client }, {});
  for await (const page of paginatedGroups) {
    const group = page.AutoScalingGroups.find(
      (g) => g.AutoScalingGroupName === groupName,
    );
    if (group) {
      return group;
    }
  }
  throw new Error(`Auto scaling group ${groupName} not found.`);
}
```

- Para obter detalhes da API, consulte os tópicos a seguir na Referência da API AWS SDK para JavaScript .
  - AttachLoadBalancerTargetGroups
  - CreateAutoScalingGroup

- CreateInstanceProfile
- CreateLaunchTemplate
- CreateListener
- CreateLoadBalancer
- CreateTargetGroup
- DeleteAutoScalingGroup
- DeleteInstanceProfile
- DeleteLaunchTemplate
- DeleteLoadBalancer
- DeleteTargetGroup
- DescribeAutoScalingGroups
- DescribeAvailabilityZones
- DescribeIamInstanceProfileAssociations
- DescribeInstances
- DescribeLoadBalancers
- DescribeSubnets
- DescribeTargetGroups
- DescribeTargetHealth
- DescribeVpcs
- RebootInstances
- ReplaceIamInstanceProfileAssociation
- TerminateInstanceInAutoScalingGroup
- UpdateAutoScalingGroup

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

Execute o cenário interativo em um prompt de comando.

```
class Runner:
    """
    Manages the deployment, demonstration, and destruction of resources for the
 resilient service.
    """

    def __init__(
        self,
        resource_path: str,
        recommendation: RecommendationService,
        autoscaler: AutoScalingWrapper,
        loadbalancer: ElasticLoadBalancerWrapper,
        param_helper: ParameterHelper,
    ):
        """
        Initializes the Runner class with the necessary parameters.

        :param resource_path: The path to resource files used by this example,
 such as IAM policies and instance scripts.
        :param recommendation: An instance of the RecommendationService class.
        :param autoscaler: An instance of the AutoScaler class.
        :param loadbalancer: An instance of the LoadBalancer class.
        :param param_helper: An instance of the ParameterHelper class.
        """
        self.resource_path = resource_path
        self.recommendation = recommendation
        self.autoscaler = autoscaler
        self.loadbalancer = loadbalancer
        self.param_helper = param_helper
        self.protocol = "HTTP"
        self.port = 80
        self.ssh_port = 22

        prefix = "doc-example-resilience"
        self.target_group_name = f"{prefix}-tg"
        self.load_balancer_name = f"{prefix}-lb"

    def deploy(self) -> None:
        """
        Deploys the resources required for the resilient service, including the
 DynamoDB table,
        EC2 instances, Auto Scaling group, and load balancer.
```

```
        """
        recommendations_path = f"{self.resource_path}/recommendations.json"
        startup_script = f"{self.resource_path}/server_startup_script.sh"
        instance_policy = f"{self.resource_path}/instance_policy.json"

        logging.info("Starting deployment of resources for the resilient
service.")

        logging.info(
            "Creating and populating DynamoDB table '%s'.",
            self.recommendation.table_name,
        )
        self.recommendation.create()
        self.recommendation.populate(recommendations_path)

        logging.info(
            "Creating an EC2 launch template with the startup script '%s'.",
            startup_script,
        )
        self.autoscaler.create_template(startup_script, instance_policy)

        logging.info(
            "Creating an EC2 Auto Scaling group across multiple Availability
Zones."
        )
        zones = self.autoscaler.create_autoscaling_group(3)

        logging.info("Creating variables that control the flow of the demo.")
        self.param_helper.reset()

        logging.info("Creating Elastic Load Balancing target group and load
balancer.")

        vpc = self.autoscaler.get_default_vpc()
        subnets = self.autoscaler.get_subnets(vpc["VpcId"], zones)
        target_group = self.loadbalancer.create_target_group(
            self.target_group_name, self.protocol, self.port, vpc["VpcId"]
        )
        self.loadbalancer.create_load_balancer(
            self.load_balancer_name, [subnet["SubnetId"] for subnet in subnets]
        )
        self.loadbalancer.create_listener(self.load_balancer_name, target_group)

        self.autoscaler.attach_load_balancer_target_group(target_group)
```

```
        logging.info("Verifying access to the load balancer endpoint.")
        endpoint = self.loadbalancer.get_endpoint(self.load_balancer_name)
        lb_success = self.loadbalancer.verify_load_balancer_endpoint(endpoint)
        current_ip_address = requests.get("http://
checkip.amazonaws.com").text.strip()

        if not lb_success:
            logging.warning(
                "Couldn't connect to the load balancer. Verifying that the port
 is open..."
            )
            sec_group, port_is_open = self.autoscaler.verify_inbound_port(
                vpc, self.port, current_ip_address
            )
            sec_group, ssh_port_is_open = self.autoscaler.verify_inbound_port(
                vpc, self.ssh_port, current_ip_address
            )
            if not port_is_open:
                logging.warning(
                    "The default security group for your VPC must allow access
 from this computer."
                )
                if q.ask(
                    f"Do you want to add a rule to security group
 {sec_group['GroupId']} to allow\n"
                    f"inbound traffic on port {self.port} from your computer's IP
 address of {current_ip_address}? (y/n) ",
                    q.is_yesno,
                ):
                    self.autoscaler.open_inbound_port(
                        sec_group["GroupId"], self.port, current_ip_address
                    )
            if not ssh_port_is_open:
                if q.ask(
                    f"Do you want to add a rule to security group
 {sec_group['GroupId']} to allow\n"
                    f"inbound SSH traffic on port {self.ssh_port} for debugging
 from your computer's IP address of {current_ip_address}? (y/n) ",
                    q.is_yesno,
                ):
                    self.autoscaler.open_inbound_port(
                        sec_group["GroupId"], self.ssh_port, current_ip_address
                    )
```

```
            lb_success =
self.loadbalancer.verify_load_balancer_endpoint(endpoint)

        if lb_success:
            logging.info(
                "Load balancer is ready. Access it at: http://%s",
current_ip_address
            )
        else:
            logging.error(
                "Couldn't get a successful response from the load balancer
endpoint. Please verify your VPC and security group settings."
            )

    def demo_choices(self) -> None:
        """
        Presents choices for interacting with the deployed service, such as
sending requests to
        the load balancer or checking the health of the targets.
        """
        actions = [
            "Send a GET request to the load balancer endpoint.",
            "Check the health of load balancer targets.",
            "Go to the next part of the demo.",
        ]
        choice = 0
        while choice != 2:
            logging.info("Choose an action to interact with the service.")
            choice = q.choose("Which action would you like to take? ", actions)
            if choice == 0:
                logging.info("Sending a GET request to the load balancer
endpoint.")
                endpoint =
self.loadbalancer.get_endpoint(self.load_balancer_name)
                logging.info("GET http://%s", endpoint)
                response = requests.get(f"http://{endpoint}")
                logging.info("Response: %s", response.status_code)
                if response.headers.get("content-type") == "application/json":
                    pp(response.json())
            elif choice == 1:
                logging.info("Checking the health of load balancer targets.")
                health =
self.loadbalancer.check_target_health(self.target_group_name)
                for target in health:
```

```
                    state = target["TargetHealth"]["State"]
                    logging.info(
                        "Target %s on port %d is %s",
                        target["Target"]["Id"],
                        target["Target"]["Port"],
                        state,
                    )
                    if state != "healthy":
                        logging.warning(
                            "%s: %s",
                            target["TargetHealth"]["Reason"],
                            target["TargetHealth"]["Description"],
                        )
                logging.info(
                    "Note that it can take a minute or two for the health check
to update."
                )
            elif choice == 2:
                logging.info("Proceeding to the next part of the demo.")

    def demo(self) -> None:
        """
        Runs the demonstration, showing how the service responds to different
failure scenarios
        and how a resilient architecture can keep the service running.
        """
        ssm_only_policy = f"{self.resource_path}/ssm_only_policy.json"

        logging.info("Resetting parameters to starting values for the demo.")
        self.param_helper.reset()

        logging.info(
            "Starting demonstration of the service's resilience under various
failure conditions."
        )
        self.demo_choices()

        logging.info(
            "Simulating failure by changing the Systems Manager parameter to a
non-existent table."
        )
        self.param_helper.put(self.param_helper.table, "this-is-not-a-table")
        logging.info("Sending GET requests will now return failure codes.")
        self.demo_choices()
```

```
        logging.info("Switching to static response mode to mitigate failure.")
        self.param_helper.put(self.param_helper.failure_response, "static")
        logging.info("Sending GET requests will now return static responses.")
        self.demo_choices()

        logging.info("Restoring normal operation of the recommendation service.")
        self.param_helper.put(self.param_helper.table,
self.recommendation.table_name)

        logging.info(
            "Introducing a failure by assigning bad credentials to one of the
instances."
        )
        self.autoscaler.create_instance_profile(
            ssm_only_policy,
            self.autoscaler.bad_creds_policy_name,
            self.autoscaler.bad_creds_role_name,
            self.autoscaler.bad_creds_profile_name,
            ["AmazonSSMManagedInstanceCore"],
        )
        instances = self.autoscaler.get_instances()
        bad_instance_id = instances[0]
        instance_profile = self.autoscaler.get_instance_profile(bad_instance_id)
        logging.info(
            "Replacing instance profile with bad credentials for instance %s.",
            bad_instance_id,
        )
        self.autoscaler.replace_instance_profile(
            bad_instance_id,
            self.autoscaler.bad_creds_profile_name,
            instance_profile["AssociationId"],
        )
        logging.info(
            "Sending GET requests may return either a valid recommendation or a
static response."
        )
        self.demo_choices()

        logging.info("Implementing deep health checks to detect unhealthy
instances.")
        self.param_helper.put(self.param_helper.health_check, "deep")
        logging.info("Checking the health of the load balancer targets.")
        self.demo_choices()
```

```
        logging.info(
            "Terminating the unhealthy instance to let the auto scaler replace
it."
        )
        self.autoscaler.terminate_instance(bad_instance_id)
        logging.info("The service remains resilient during instance
replacement.")
        self.demo_choices()

        logging.info("Simulating a complete failure of the recommendation
service.")
        self.param_helper.put(self.param_helper.table, "this-is-not-a-table")
        logging.info(
            "All instances will report as unhealthy, but the service will still
return static responses."
        )
        self.demo_choices()
        self.param_helper.reset()

    def destroy(self, automation=False) -> None:
        """
        Destroys all resources created for the demo, including the load balancer,
Auto Scaling group,
        EC2 instances, and DynamoDB table.
        """
        logging.info(
            "This concludes the demo. Preparing to clean up all AWS resources
created during the demo."
        )
        if automation:
            cleanup = True
        else:
            cleanup = q.ask(
                "Do you want to clean up all demo resources? (y/n) ", q.is_yesno
            )

        if cleanup:
            logging.info("Deleting load balancer and related resources.")
            self.loadbalancer.delete_load_balancer(self.load_balancer_name)
            self.loadbalancer.delete_target_group(self.target_group_name)
            self.autoscaler.delete_autoscaling_group(self.autoscaler.group_name)
            self.autoscaler.delete_key_pair()
            self.autoscaler.delete_template()
```

```
            self.autoscaler.delete_instance_profile(
                self.autoscaler.bad_creds_profile_name,
                self.autoscaler.bad_creds_role_name,
            )
            logging.info("Deleting DynamoDB table and other resources.")
            self.recommendation.destroy()
        else:
            logging.warning(
                "Resources have not been deleted. Ensure you clean them up
 manually to avoid unexpected charges."
            )


def main() -> None:
    """
    Main function to parse arguments and run the appropriate actions for the
 demo.
    """
    parser = argparse.ArgumentParser()
    parser.add_argument(
        "--action",
        required=True,
        choices=["all", "deploy", "demo", "destroy"],
        help="The action to take for the demo. When 'all' is specified, resources
 are\n"
        "deployed, the demo is run, and resources are destroyed.",
    )
    parser.add_argument(
        "--resource_path",
        default="../../../scenarios/features/resilient_service/resources",
        help="The path to resource files used by this example, such as IAM
 policies and\n"
        "instance scripts.",
    )
    args = parser.parse_args()

    logging.info("Starting the Resilient Service demo.")

    prefix = "doc-example-resilience"

    # Service Clients
    ddb_client = boto3.client("dynamodb")
    elb_client = boto3.client("elbv2")
    autoscaling_client = boto3.client("autoscaling")
```

```
    ec2_client = boto3.client("ec2")
    ssm_client = boto3.client("ssm")
    iam_client = boto3.client("iam")

    # Wrapper instantiations
    recommendation = RecommendationService(
        "doc-example-recommendation-service", ddb_client
    )
    autoscaling_wrapper = AutoScalingWrapper(
        prefix,
        "t3.micro",
        "/aws/service/ami-amazon-linux-latest/amzn2-ami-hvm-x86_64-gp2",
        autoscaling_client,
        ec2_client,
        ssm_client,
        iam_client,
    )
    elb_wrapper = ElasticLoadBalancerWrapper(elb_client)
    param_helper = ParameterHelper(recommendation.table_name, ssm_client)

    # Demo invocation
    runner = Runner(
        args.resource_path,
        recommendation,
        autoscaling_wrapper,
        elb_wrapper,
        param_helper,
    )
    actions = [args.action] if args.action != "all" else ["deploy", "demo",
 "destroy"]
    for action in actions:
        if action == "deploy":
            runner.deploy()
        elif action == "demo":
            runner.demo()
        elif action == "destroy":
            runner.destroy()

    logging.info("Demo completed successfully.")


if __name__ == "__main__":
    logging.basicConfig(level=logging.INFO, format="%(levelname)s: %(message)s")
    main()
```

Crie uma classe que envolva ações do Auto Scaling e do Amazon EC2.

```
class AutoScalingWrapper:
    """
    Encapsulates Amazon EC2 Auto Scaling and EC2 management actions.
    """

    def __init__(
        self,
        resource_prefix: str,
        inst_type: str,
        ami_param: str,
        autoscaling_client: boto3.client,
        ec2_client: boto3.client,
        ssm_client: boto3.client,
        iam_client: boto3.client,
    ):
        """
        Initializes the AutoScaler class with the necessary parameters.

        :param resource_prefix: The prefix for naming AWS resources that are
 created by this class.
        :param inst_type: The type of EC2 instance to create, such as t3.micro.
        :param ami_param: The Systems Manager parameter used to look up the AMI
 that is created.
        :param autoscaling_client: A Boto3 EC2 Auto Scaling client.
        :param ec2_client: A Boto3 EC2 client.
        :param ssm_client: A Boto3 Systems Manager client.
        :param iam_client: A Boto3 IAM client.
        """
        self.inst_type = inst_type
        self.ami_param = ami_param
        self.autoscaling_client = autoscaling_client
        self.ec2_client = ec2_client
        self.ssm_client = ssm_client
        self.iam_client = iam_client
        sts_client = boto3.client("sts")
        self.account_id = sts_client.get_caller_identity()["Account"]

        self.key_pair_name = f"{resource_prefix}-key-pair"
        self.launch_template_name = f"{resource_prefix}-template-"
```

```
        self.group_name = f"{resource_prefix}-group"

        # Happy path
        self.instance_policy_name = f"{resource_prefix}-pol"
        self.instance_role_name = f"{resource_prefix}-role"
        self.instance_profile_name = f"{resource_prefix}-prof"

        # Failure mode
        self.bad_creds_policy_name = f"{resource_prefix}-bc-pol"
        self.bad_creds_role_name = f"{resource_prefix}-bc-role"
        self.bad_creds_profile_name = f"{resource_prefix}-bc-prof"


    def create_policy(self, policy_file: str, policy_name: str) -> str:
        """
        Creates a new IAM policy or retrieves the ARN of an existing policy.

        :param policy_file: The path to a JSON file that contains the policy
 definition.
        :param policy_name: The name to give the created policy.
        :return: The ARN of the created or existing policy.
        """
        with open(policy_file) as file:
            policy_doc = file.read()

        try:
            response = self.iam_client.create_policy(
                PolicyName=policy_name, PolicyDocument=policy_doc
            )
            policy_arn = response["Policy"]["Arn"]
            log.info(f"Policy '{policy_name}' created successfully. ARN:
 {policy_arn}")
            return policy_arn

        except ClientError as err:
            if err.response["Error"]["Code"] == "EntityAlreadyExists":
                # If the policy already exists, get its ARN
                response = self.iam_client.get_policy(
                    PolicyArn=f"arn:aws:iam::{self.account_id}:policy/
{policy_name}"
                )
                policy_arn = response["Policy"]["Arn"]
                log.info(f"Policy '{policy_name}' already exists. ARN:
 {policy_arn}")
```

```
                return policy_arn
            log.error(f"Full error:\n\t{err}")

    def create_role(self, role_name: str, assume_role_doc: dict) -> str:
        """
        Creates a new IAM role or retrieves the ARN of an existing role.

        :param role_name: The name to give the created role.
        :param assume_role_doc: The assume role policy document that specifies
 which
                                entities can assume the role.
        :return: The ARN of the created or existing role.
        """
        try:
            response = self.iam_client.create_role(
                RoleName=role_name,
 AssumeRolePolicyDocument=json.dumps(assume_role_doc)
            )
            role_arn = response["Role"]["Arn"]
            log.info(f"Role '{role_name}' created successfully. ARN: {role_arn}")
            return role_arn

        except ClientError as err:
            if err.response["Error"]["Code"] == "EntityAlreadyExists":
                # If the role already exists, get its ARN
                response = self.iam_client.get_role(RoleName=role_name)
                role_arn = response["Role"]["Arn"]
                log.info(f"Role '{role_name}' already exists. ARN: {role_arn}")
                return role_arn
            log.error(f"Full error:\n\t{err}")

    def attach_policy(
        self,
        role_name: str,
        policy_arn: str,
        aws_managed_policies: Tuple[str, ...] = (),
    ) -> None:
        """
        Attaches an IAM policy to a role and optionally attaches additional AWS-
managed policies.

        :param role_name: The name of the role to attach the policy to.
        :param policy_arn: The ARN of the policy to attach.
```

```
        :param aws_managed_policies: A tuple of AWS-managed policy names to
attach to the role.
        """
        try:
            self.iam_client.attach_role_policy(RoleName=role_name,
PolicyArn=policy_arn)
            for aws_policy in aws_managed_policies:
                self.iam_client.attach_role_policy(
                    RoleName=role_name,
                    PolicyArn=f"arn:aws:iam::aws:policy/{aws_policy}",
                )
            log.info(f"Attached policy {policy_arn} to role {role_name}.")
        except ClientError as err:
            log.error(f"Failed to attach policy {policy_arn} to role
{role_name}.")
            log.error(f"Full error:\n\t{err}")

    def create_instance_profile(
        self,
        policy_file: str,
        policy_name: str,
        role_name: str,
        profile_name: str,
        aws_managed_policies: Tuple[str, ...] = (),
    ) -> str:
        """
        Creates a policy, role, and profile that is associated with instances
created by
        this class. An instance's associated profile defines a role that is
assumed by the
        instance. The role has attached policies that specify the AWS permissions
granted to
        clients that run on the instance.

        :param policy_file: The name of a JSON file that contains the policy
definition to
                            create and attach to the role.
        :param policy_name: The name to give the created policy.
        :param role_name: The name to give the created role.
        :param profile_name: The name to the created profile.
        :param aws_managed_policies: Additional AWS-managed policies that are
attached to
                                     the role, such as
AmazonSSMManagedInstanceCore to grant
```

```
                                    use of Systems Manager to send commands to
the instance.
        :return: The ARN of the profile that is created.
        """
        assume_role_doc = {
            "Version":"2012-10-17",
            "Statement": [
                {
                    "Effect": "Allow",
                    "Principal": {"Service": "ec2.amazonaws.com"},
                    "Action": "sts:AssumeRole",
                }
            ],
        }
        policy_arn = self.create_policy(policy_file, policy_name)
        self.create_role(role_name, assume_role_doc)
        self.attach_policy(role_name, policy_arn, aws_managed_policies)

        try:
            profile_response = self.iam_client.create_instance_profile(
                InstanceProfileName=profile_name
            )
            waiter = self.iam_client.get_waiter("instance_profile_exists")
            waiter.wait(InstanceProfileName=profile_name)
            time.sleep(10)  # wait a little longer
            profile_arn = profile_response["InstanceProfile"]["Arn"]
            self.iam_client.add_role_to_instance_profile(
                InstanceProfileName=profile_name, RoleName=role_name
            )
            log.info("Created profile %s and added role %s.", profile_name,
role_name)
        except ClientError as err:
            if err.response["Error"]["Code"] == "EntityAlreadyExists":
                prof_response = self.iam_client.get_instance_profile(
                    InstanceProfileName=profile_name
                )
                profile_arn = prof_response["InstanceProfile"]["Arn"]
                log.info(
                    "Instance profile %s already exists, nothing to do.",
profile_name
                )
            log.error(f"Full error:\n\t{err}")
        return profile_arn
```

```
   def get_instance_profile(self, instance_id: str) -> Dict[str, Any]:
       """
       Gets data about the profile associated with an instance.

       :param instance_id: The ID of the instance to look up.
       :return: The profile data.
       """
       try:
           response =
self.ec2_client.describe_iam_instance_profile_associations(
               Filters=[{"Name": "instance-id", "Values": [instance_id]}]
           )
           if not response["IamInstanceProfileAssociations"]:
               log.info(f"No instance profile found for instance
{instance_id}.")
           profile_data = response["IamInstanceProfileAssociations"][0]
           log.info(f"Retrieved instance profile for instance {instance_id}.")
           return profile_data
       except ClientError as err:
           log.error(
               f"Failed to retrieve instance profile for instance
{instance_id}."
           )
           error_code = err.response["Error"]["Code"]
           if error_code == "InvalidInstanceID.NotFound":
               log.error(f"The instance ID '{instance_id}' does not exist.")
           log.error(f"Full error:\n\t{err}")


   def replace_instance_profile(
       self,
       instance_id: str,
       new_instance_profile_name: str,
       profile_association_id: str,
   ) -> None:
       """
       Replaces the profile associated with a running instance. After the
profile is
       replaced, the instance is rebooted to ensure that it uses the new
profile. When
       the instance is ready, Systems Manager is used to restart the Python web
server.
```

```
        :param instance_id: The ID of the instance to restart.
        :param new_instance_profile_name: The name of the new profile to
associate with
                                          the specified instance.
        :param profile_association_id: The ID of the existing profile association
for the
                                       instance.
        """
        try:
            self.ec2_client.replace_iam_instance_profile_association(
                IamInstanceProfile={"Name": new_instance_profile_name},
                AssociationId=profile_association_id,
            )
            log.info(
                "Replaced instance profile for association %s with profile %s.",
                profile_association_id,
                new_instance_profile_name,
            )
            time.sleep(5)

            self.ec2_client.reboot_instances(InstanceIds=[instance_id])
            log.info("Rebooting instance %s.", instance_id)
            waiter = self.ec2_client.get_waiter("instance_running")
            log.info("Waiting for instance %s to be running.", instance_id)
            waiter.wait(InstanceIds=[instance_id])
            log.info("Instance %s is now running.", instance_id)

            self.ssm_client.send_command(
                InstanceIds=[instance_id],
                DocumentName="AWS-RunShellScript",
                Parameters={"commands": ["cd / && sudo python3 server.py 80"]},
            )
            log.info(f"Restarted the Python web server on instance
'{instance_id}'.")
        except ClientError as err:
            log.error("Failed to replace instance profile.")
            error_code = err.response["Error"]["Code"]
            if error_code == "InvalidAssociationID.NotFound":
                log.error(
                    f"Association ID '{profile_association_id}' does not exist."
                    "Please check the association ID and try again."
                )
            if error_code == "InvalidInstanceId":
                log.error(
```

```
                    f"The specified instance ID '{instance_id}' does not exist or
is not available for SSM. "
                    f"Please verify the instance ID and try again."
                )
            log.error(f"Full error:\n\t{err}")


    def delete_instance_profile(self, profile_name: str, role_name: str) -> None:
        """
        Detaches a role from an instance profile, detaches policies from the
role,
        and deletes all the resources.

        :param profile_name: The name of the profile to delete.
        :param role_name: The name of the role to delete.
        """
        try:
            self.iam_client.remove_role_from_instance_profile(
                InstanceProfileName=profile_name, RoleName=role_name
            )

self.iam_client.delete_instance_profile(InstanceProfileName=profile_name)
            log.info("Deleted instance profile %s.", profile_name)
            attached_policies = self.iam_client.list_attached_role_policies(
                RoleName=role_name
            )
            for pol in attached_policies["AttachedPolicies"]:
                self.iam_client.detach_role_policy(
                    RoleName=role_name, PolicyArn=pol["PolicyArn"]
                )
                if not pol["PolicyArn"].startswith("arn:aws:iam::aws"):
                    self.iam_client.delete_policy(PolicyArn=pol["PolicyArn"])
                log.info("Detached and deleted policy %s.", pol["PolicyName"])
            self.iam_client.delete_role(RoleName=role_name)
            log.info("Deleted role %s.", role_name)
        except ClientError as err:
            log.error(
                f"Couldn't delete instance profile {profile_name} or detach "
                f"policies and delete role {role_name}: {err}"
            )
            if err.response["Error"]["Code"] == "NoSuchEntity":
                log.info(
                    "Instance profile %s doesn't exist, nothing to do.",
profile_name
```

```
                )


    def create_key_pair(self, key_pair_name: str) -> None:
        """
        Creates a new key pair.

        :param key_pair_name: The name of the key pair to create.
        """
        try:
            response = self.ec2_client.create_key_pair(KeyName=key_pair_name)
            with open(f"{key_pair_name}.pem", "w") as file:
                file.write(response["KeyMaterial"])
            chmod(f"{key_pair_name}.pem", 0o600)
            log.info("Created key pair %s.", key_pair_name)
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(f"Failed to create key pair {key_pair_name}.")
            if error_code == "InvalidKeyPair.Duplicate":
                log.error(f"A key pair with the name '{key_pair_name}' already
exists.")
            log.error(f"Full error:\n\t{err}")


    def delete_key_pair(self) -> None:
        """
        Deletes a key pair.
        """
        try:
            self.ec2_client.delete_key_pair(KeyName=self.key_pair_name)
            remove(f"{self.key_pair_name}.pem")
            log.info("Deleted key pair %s.", self.key_pair_name)
        except ClientError as err:
            log.error(f"Couldn't delete key pair '{self.key_pair_name}'.")
            log.error(f"Full error:\n\t{err}")
        except FileNotFoundError as err:
            log.info("Key pair %s doesn't exist, nothing to do.",
self.key_pair_name)
            log.error(f"Full error:\n\t{err}")


    def create_template(
        self, server_startup_script_file: str, instance_policy_file: str
    ) -> Dict[str, Any]:
```

```
        """
        Creates an Amazon EC2 launch template to use with Amazon EC2 Auto
Scaling. The
        launch template specifies a Bash script in its user data field that runs
after
        the instance is started. This script installs Python packages and starts
a
        Python web server on the instance.

        :param server_startup_script_file: The path to a Bash script file that is
run
                                           when an instance starts.
        :param instance_policy_file: The path to a file that defines a
permissions policy
                                     to create and attach to the instance
profile.
        :return: Information about the newly created template.
        """
        template = {}
        try:
            # Create key pair and instance profile
            self.create_key_pair(self.key_pair_name)
            self.create_instance_profile(
                instance_policy_file,
                self.instance_policy_name,
                self.instance_role_name,
                self.instance_profile_name,
            )

            # Read the startup script
            with open(server_startup_script_file) as file:
                start_server_script = file.read()

            # Get the latest AMI ID
            ami_latest = self.ssm_client.get_parameter(Name=self.ami_param)
            ami_id = ami_latest["Parameter"]["Value"]

            # Create the launch template
            lt_response = self.ec2_client.create_launch_template(
                LaunchTemplateName=self.launch_template_name,
                LaunchTemplateData={
                    "InstanceType": self.inst_type,
                    "ImageId": ami_id,
                    "IamInstanceProfile": {"Name": self.instance_profile_name},
```

```
                    "UserData": base64.b64encode(
                        start_server_script.encode(encoding="utf-8")
                    ).decode(encoding="utf-8"),
                    "KeyName": self.key_pair_name,
                },
            )
            template = lt_response["LaunchTemplate"]
            log.info(
                f"Created launch template {self.launch_template_name} for AMI
{ami_id} on {self.inst_type}."
            )
        except ClientError as err:
            log.error(f"Failed to create launch template
{self.launch_template_name}.")
            error_code = err.response["Error"]["Code"]
            if error_code == "InvalidLaunchTemplateName.AlreadyExistsException":
                log.info(
                    f"Launch template {self.launch_template_name} already exists,
nothing to do."
                )
            log.error(f"Full error:\n\t{err}")
        return template


    def delete_template(self):
        """
        Deletes a launch template.
        """
        try:
            self.ec2_client.delete_launch_template(
                LaunchTemplateName=self.launch_template_name
            )
            self.delete_instance_profile(
                self.instance_profile_name, self.instance_role_name
            )
            log.info("Launch template %s deleted.", self.launch_template_name)
        except ClientError as err:
            if (
                err.response["Error"]["Code"]
                == "InvalidLaunchTemplateName.NotFoundException"
            ):
                log.info(
                    "Launch template %s does not exist, nothing to do.",
                    self.launch_template_name,
```

```
                )
            log.error(f"Full error:\n\t{err}")


    def get_availability_zones(self) -> List[str]:
        """
        Gets a list of Availability Zones in the AWS Region of the Amazon EC2
client.

        :return: The list of Availability Zones for the client Region.
        """
        try:
            response = self.ec2_client.describe_availability_zones()
            zones = [zone["ZoneName"] for zone in response["AvailabilityZones"]]
            log.info(f"Retrieved {len(zones)} availability zones: {zones}.")
        except ClientError as err:
            log.error("Failed to retrieve availability zones.")
            log.error(f"Full error:\n\t{err}")
        else:
            return zones


    def create_autoscaling_group(self, group_size: int) -> List[str]:
        """
        Creates an EC2 Auto Scaling group with the specified size.

        :param group_size: The number of instances to set for the minimum and
maximum in
                           the group.
        :return: The list of Availability Zones specified for the group.
        """
        try:
            zones = self.get_availability_zones()
            self.autoscaling_client.create_auto_scaling_group(
                AutoScalingGroupName=self.group_name,
                AvailabilityZones=zones,
                LaunchTemplate={
                    "LaunchTemplateName": self.launch_template_name,
                    "Version": "$Default",
                },
                MinSize=group_size,
                MaxSize=group_size,
            )
            log.info(
```

```
                f"Created EC2 Auto Scaling group {self.group_name} with
availability zones {zones}."
            )
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            if error_code == "AlreadyExists":
                log.info(
                    f"EC2 Auto Scaling group {self.group_name} already exists,
nothing to do."
                )
            else:
                log.error(f"Failed to create EC2 Auto Scaling group
{self.group_name}.")
                log.error(f"Full error:\n\t{err}")
        else:
            return zones


    def get_instances(self) -> List[str]:
        """
        Gets data about the instances in the EC2 Auto Scaling group.

        :return: A list of instance IDs in the Auto Scaling group.
        """
        try:
            as_response = self.autoscaling_client.describe_auto_scaling_groups(
                AutoScalingGroupNames=[self.group_name]
            )
            instance_ids = [
                i["InstanceId"]
                for i in as_response["AutoScalingGroups"][0]["Instances"]
            ]
            log.info(
                f"Retrieved {len(instance_ids)} instances for Auto Scaling group
{self.group_name}."
            )
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(
                f"Failed to retrieve instances for Auto Scaling group
{self.group_name}."
            )
            if error_code == "ResourceNotFound":
```

```
                log.error(f"The Auto Scaling group '{self.group_name}' does not
exist.")
            log.error(f"Full error:\n\t{err}")
        else:
            return instance_ids


    def terminate_instance(self, instance_id: str, decrementsetting=False) ->
None:
        """
        Terminates an instance in an EC2 Auto Scaling group. After an instance is
        terminated, it can no longer be accessed.

        :param instance_id: The ID of the instance to terminate.
        :param decrementsetting: If True, do not replace terminated instances.
        """
        try:
            self.autoscaling_client.terminate_instance_in_auto_scaling_group(
                InstanceId=instance_id,
                ShouldDecrementDesiredCapacity=decrementsetting,
            )
            log.info("Terminated instance %s.", instance_id)

            # Adding a waiter to ensure the instance is terminated
            waiter = self.ec2_client.get_waiter("instance_terminated")
            log.info("Waiting for instance %s to be terminated...", instance_id)
            waiter.wait(InstanceIds=[instance_id])
            log.info(
                f"Instance '{instance_id}' has been terminated and will be
replaced."
            )

        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(f"Failed to terminate instance '{instance_id}'.")
            if error_code == "ScalingActivityInProgressFault":
                log.error(
                    "Scaling activity is currently in progress. "
                    "Wait for the scaling activity to complete before attempting
to terminate the instance again."
                )
            elif error_code == "ResourceContentionFault":
                log.error(
                    "The request failed due to a resource contention issue. "
```

```
                    "Ensure that no conflicting operations are being performed on
the resource."
                )
            log.error(f"Full error:\n\t{err}")

    def attach_load_balancer_target_group(
        self, lb_target_group: Dict[str, Any]
    ) -> None:
        """
        Attaches an Elastic Load Balancing (ELB) target group to this EC2 Auto
Scaling group.
        The target group specifies how the load balancer forwards requests to the
instances
        in the group.

        :param lb_target_group: Data about the ELB target group to attach.
        """
        try:
            self.autoscaling_client.attach_load_balancer_target_groups(
                AutoScalingGroupName=self.group_name,
                TargetGroupARNs=[lb_target_group["TargetGroupArn"]],
            )
            log.info(
                "Attached load balancer target group %s to auto scaling group
%s.",
                lb_target_group["TargetGroupName"],
                self.group_name,
            )
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(
                f"Failed to attach load balancer target group
'{lb_target_group['TargetGroupName']}'."
            )
            if error_code == "ResourceContentionFault":
                log.error(
                    "The request failed due to a resource contention issue. "
                    "Ensure that no conflicting operations are being performed on
the resource."
                )
            elif error_code == "ServiceLinkedRoleFailure":
                log.error(
                    "The operation failed because the service-linked role is not
ready or does not exist. "
```

```
                    "Check that the service-linked role exists and is correctly
 configured."
                )
            log.error(f"Full error:\n\t{err}")


    def delete_autoscaling_group(self, group_name: str) -> None:
        """
        Terminates all instances in the group, then deletes the EC2 Auto Scaling
 group.

        :param group_name: The name of the group to delete.
        """
        try:
            response = self.autoscaling_client.describe_auto_scaling_groups(
                AutoScalingGroupNames=[group_name]
            )
            groups = response.get("AutoScalingGroups", [])
            if len(groups) > 0:
                self.autoscaling_client.update_auto_scaling_group(
                    AutoScalingGroupName=group_name, MinSize=0
                )
                instance_ids = [inst["InstanceId"] for inst in groups[0]
["Instances"]]
                for inst_id in instance_ids:
                    self.terminate_instance(inst_id)

                # Wait for all instances to be terminated
                if instance_ids:
                    waiter = self.ec2_client.get_waiter("instance_terminated")
                    log.info("Waiting for all instances to be terminated...")
                    waiter.wait(InstanceIds=instance_ids)
                    log.info("All instances have been terminated.")
            else:
                log.info(f"No groups found named '{group_name}'! Nothing to do.")
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(f"Failed to delete Auto Scaling group '{group_name}'.")
            if error_code == "ScalingActivityInProgressFault":
                log.error(
                    "Scaling activity is currently in progress. "
                    "Wait for the scaling activity to complete before attempting
 to delete the group again."
                )
```

```
            elif error_code == "ResourceContentionFault":
                log.error(
                    "The request failed due to a resource contention issue. "
                    "Ensure that no conflicting operations are being performed on
 the group."
                )
            log.error(f"Full error:\n\t{err}")


    def get_default_vpc(self) -> Dict[str, Any]:
        """
        Gets the default VPC for the account.

        :return: Data about the default VPC.
        """
        try:
            response = self.ec2_client.describe_vpcs(
                Filters=[{"Name": "is-default", "Values": ["true"]}]
            )
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error("Failed to retrieve the default VPC.")
            if error_code == "UnauthorizedOperation":
                log.error(
                    "You do not have the necessary permissions to describe VPCs.
 "
                    "Ensure that your AWS IAM user or role has the correct
 permissions."
                )
            elif error_code == "InvalidParameterValue":
                log.error(
                    "One or more parameters are invalid. Check the request
 parameters."
                )

            log.error(f"Full error:\n\t{err}")
        else:
            if "Vpcs" in response and response["Vpcs"]:
                log.info(f"Retrieved default VPC: {response['Vpcs'][0]
['VpcId']}")
                return response["Vpcs"][0]
            else:
                pass
```

```
   def verify_inbound_port(
       self, vpc: Dict[str, Any], port: int, ip_address: str
   ) -> Tuple[Dict[str, Any], bool]:
       """
       Verify the default security group of the specified VPC allows ingress
from this
       computer. This can be done by allowing ingress from this computer's IP
       address. In some situations, such as connecting from a corporate network,
you
       must instead specify a prefix list ID. You can also temporarily open the
port to
       any IP address while running this example. If you do, be sure to remove
public
       access when you're done.

       :param vpc: The VPC used by this example.
       :param port: The port to verify.
       :param ip_address: This computer's IP address.
       :return: The default security group of the specified VPC, and a value
that indicates
                whether the specified port is open.
       """
       try:
           response = self.ec2_client.describe_security_groups(
               Filters=[
                   {"Name": "group-name", "Values": ["default"]},
                   {"Name": "vpc-id", "Values": [vpc["VpcId"]]},
               ]
           )
           sec_group = response["SecurityGroups"][0]
           port_is_open = False
           log.info(f"Found default security group {sec_group['GroupId']}.")

           for ip_perm in sec_group["IpPermissions"]:
               if ip_perm.get("FromPort", 0) == port:
                   log.info(f"Found inbound rule: {ip_perm}")
                   for ip_range in ip_perm["IpRanges"]:
                       cidr = ip_range.get("CidrIp", "")
                       if cidr.startswith(ip_address) or cidr == "0.0.0.0/0":
                           port_is_open = True
                   if ip_perm["PrefixListIds"]:
                       port_is_open = True
                   if not port_is_open:
```

```
                    log.info(
                        f"The inbound rule does not appear to be open to
either this computer's IP "
                        f"address of {ip_address}, to all IP addresses
(0.0.0.0/0), or to a prefix list ID."
                    )
                else:
                    break
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(
                f"Failed to verify inbound rule for port {port} for VPC
{vpc['VpcId']}."
            )
            if error_code == "InvalidVpcID.NotFound":
                log.error(
                    f"The specified VPC ID '{vpc['VpcId']}' does not exist.
Please check the VPC ID."
                )
            log.error(f"Full error:\n\t{err}")
        else:
            return sec_group, port_is_open


    def open_inbound_port(self, sec_group_id: str, port: int, ip_address: str) ->
None:
        """
        Add an ingress rule to the specified security group that allows access on
the
        specified port from the specified IP address.

        :param sec_group_id: The ID of the security group to modify.
        :param port: The port to open.
        :param ip_address: The IP address that is granted access.
        """
        try:
            self.ec2_client.authorize_security_group_ingress(
                GroupId=sec_group_id,
                CidrIp=f"{ip_address}/32",
                FromPort=port,
                ToPort=port,
                IpProtocol="tcp",
            )
            log.info(
```

```
                "Authorized ingress to %s on port %s from %s.",
                sec_group_id,
                port,
                ip_address,
            )
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(
                f"Failed to authorize ingress to security group '{sec_group_id}'
on port {port} from {ip_address}."
            )
            if error_code == "InvalidGroupId.Malformed":
                log.error(
                    "The security group ID is malformed. "
                    "Please verify that the security group ID is correct."
                )
            elif error_code == "InvalidPermission.Duplicate":
                log.error(
                    "The specified rule already exists in the security group. "
                    "Check the existing rules for this security group."
                )
            log.error(f"Full error:\n\t{err}")


    def get_subnets(self, vpc_id: str, zones: List[str] = None) -> List[Dict[str,
Any]]:
        """
        Gets the default subnets in a VPC for a specified list of Availability
Zones.

        :param vpc_id: The ID of the VPC to look up.
        :param zones: The list of Availability Zones to look up.
        :return: The list of subnets found.
        """
        # Ensure that 'zones' is a list, even if None is passed
        if zones is None:
            zones = []
        try:
            paginator = self.ec2_client.get_paginator("describe_subnets")
            page_iterator = paginator.paginate(
                Filters=[
                    {"Name": "vpc-id", "Values": [vpc_id]},
                    {"Name": "availability-zone", "Values": zones},
                    {"Name": "default-for-az", "Values": ["true"]},
```

```
                ]
            )

            subnets = []
            for page in page_iterator:
                subnets.extend(page["Subnets"])

            log.info("Found %s subnets for the specified zones.", len(subnets))
            return subnets
        except ClientError as err:
            log.error(
                f"Failed to retrieve subnets for VPC '{vpc_id}' in zones
 {zones}."
            )
            error_code = err.response["Error"]["Code"]
            if error_code == "InvalidVpcID.NotFound":
                log.error(
                    "The specified VPC ID does not exist. "
                    "Please check the VPC ID and try again."
                )
            # Add more error-specific handling as needed
            log.error(f"Full error:\n\t{err}")
```

Crie uma classe que envolva ações do Elastic Load Balancing.

```
class ElasticLoadBalancerWrapper:
    """Encapsulates Elastic Load Balancing (ELB) actions."""

    def __init__(self, elb_client: boto3.client):
        """
        Initializes the LoadBalancer class with the necessary parameters.
        """
        self.elb_client = elb_client


    def create_target_group(
        self, target_group_name: str, protocol: str, port: int, vpc_id: str
    ) -> Dict[str, Any]:
        """
```

```
        Creates an Elastic Load Balancing target group. The target group
specifies how
        the load balancer forwards requests to instances in the group and how
instance
        health is checked.

        To speed up this demo, the health check is configured with shortened
times and
        lower thresholds. In production, you might want to decrease the
sensitivity of
        your health checks to avoid unwanted failures.

        :param target_group_name: The name of the target group to create.
        :param protocol: The protocol to use to forward requests, such as 'HTTP'.
        :param port: The port to use to forward requests, such as 80.
        :param vpc_id: The ID of the VPC in which the load balancer exists.
        :return: Data about the newly created target group.
        """
        try:
            response = self.elb_client.create_target_group(
                Name=target_group_name,
                Protocol=protocol,
                Port=port,
                HealthCheckPath="/healthcheck",
                HealthCheckIntervalSeconds=10,
                HealthCheckTimeoutSeconds=5,
                HealthyThresholdCount=2,
                UnhealthyThresholdCount=2,
                VpcId=vpc_id,
            )
            target_group = response["TargetGroups"][0]
            log.info(f"Created load balancing target group
'{target_group_name}'.")
            return target_group
        except ClientError as err:
            log.error(
                f"Couldn't create load balancing target group
'{target_group_name}'."
            )
            error_code = err.response["Error"]["Code"]

            if error_code == "DuplicateTargetGroupName":
                log.error(
                    f"Target group name {target_group_name} already exists. "
```

```
                    "Check if the target group already exists."
                    "Consider using a different name or deleting the existing
target group if appropriate."
                )
            elif error_code == "TooManyTargetGroups":
                log.error(
                    "Too many target groups exist in the account. "
                    "Consider deleting unused target groups to create space for
new ones."
                )
            log.error(f"Full error:\n\t{err}")


    def delete_target_group(self, target_group_name) -> None:
        """
        Deletes the target group.
        """
        try:
            # Describe the target group to get its ARN
            response =
self.elb_client.describe_target_groups(Names=[target_group_name])
            tg_arn = response["TargetGroups"][0]["TargetGroupArn"]

            # Delete the target group
            self.elb_client.delete_target_group(TargetGroupArn=tg_arn)
            log.info("Deleted load balancing target group %s.",
target_group_name)

            # Use a custom waiter to wait until the target group is no longer
available
            self.wait_for_target_group_deletion(self.elb_client, tg_arn)
            log.info("Target group %s successfully deleted.", target_group_name)

        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(f"Failed to delete target group '{target_group_name}'.")
            if error_code == "TargetGroupNotFound":
                log.error(
                    "Load balancer target group either already deleted or never
existed. "
                    "Verify the name and check that the resource exists in the
AWS Console."
                )
            elif error_code == "ResourceInUseException":
```

```
                log.error(
                    "Target group still in use by another resource. "
                    "Ensure that the target group is no longer associated with
any load balancers or resources.",
                )
            log.error(f"Full error:\n\t{err}")

    def wait_for_target_group_deletion(
        self, elb_client, target_group_arn, max_attempts=10, delay=30
    ):
        for attempt in range(max_attempts):
            try:

elb_client.describe_target_groups(TargetGroupArns=[target_group_arn])
                print(
                    f"Attempt {attempt + 1}: Target group {target_group_arn}
still exists."
                )
            except ClientError as e:
                if e.response["Error"]["Code"] == "TargetGroupNotFound":
                    print(
                        f"Target group {target_group_arn} has been successfully
deleted."
                    )
                    return
                else:
                    raise
            time.sleep(delay)
        raise TimeoutError(
            f"Target group {target_group_arn} was not deleted after {max_attempts
* delay} seconds."
        )


    def create_load_balancer(
        self,
        load_balancer_name: str,
        subnet_ids: List[str],
    ) -> Dict[str, Any]:
        """
        Creates an Elastic Load Balancing load balancer that uses the specified
subnets
        and forwards requests to the specified target group.
```

```
        :param load_balancer_name: The name of the load balancer to create.
        :param subnet_ids: A list of subnets to associate with the load balancer.
        :return: Data about the newly created load balancer.
        """
        try:
            response = self.elb_client.create_load_balancer(
                Name=load_balancer_name, Subnets=subnet_ids
            )
            load_balancer = response["LoadBalancers"][0]
            log.info(f"Created load balancer '{load_balancer_name}'.")

            waiter = self.elb_client.get_waiter("load_balancer_available")
            log.info(
                f"Waiting for load balancer '{load_balancer_name}' to be
available..."
            )
            waiter.wait(Names=[load_balancer_name])
            log.info(f"Load balancer '{load_balancer_name}' is now available!")

        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(
                f"Failed to create load balancer '{load_balancer_name}'. Error
code: {error_code}, Message: {err.response['Error']['Message']}"
            )

            if error_code == "DuplicateLoadBalancerNameException":
                log.error(
                    f"A load balancer with the name '{load_balancer_name}'
already exists. "
                    "Load balancer names must be unique within the AWS region. "
                    "Please choose a different name and try again."
                )
            if error_code == "TooManyLoadBalancersException":
                log.error(
                    "The maximum number of load balancers has been reached in
this account and region. "
                    "You can delete unused load balancers or request an increase
in the service quota from AWS Support."
                )
            log.error(f"Full error:\n\t{err}")
        else:
            return load_balancer
```

```
   def create_listener(
       self,
       load_balancer_name: str,
       target_group: Dict[str, Any],
   ) -> Dict[str, Any]:
       """
       Creates a listener for the specified load balancer that forwards requests
to the
       specified target group.

       :param load_balancer_name: The name of the load balancer to create a
listener for.
       :param target_group: An existing target group that is added as a listener
to the
                            load balancer.
       :return: Data about the newly created listener.
       """
       try:
           # Retrieve the load balancer ARN
           load_balancer_response = self.elb_client.describe_load_balancers(
               Names=[load_balancer_name]
           )
           load_balancer_arn = load_balancer_response["LoadBalancers"][0][
               "LoadBalancerArn"
           ]

           # Create the listener
           response = self.elb_client.create_listener(
               LoadBalancerArn=load_balancer_arn,
               Protocol=target_group["Protocol"],
               Port=target_group["Port"],
               DefaultActions=[
                   {
                       "Type": "forward",
                       "TargetGroupArn": target_group["TargetGroupArn"],
                   }
               ],
           )
           log.info(
               f"Created listener to forward traffic from load balancer
'{load_balancer_name}' to target group '{target_group['TargetGroupName']}'."
           )
           return response["Listeners"][0]
```

```
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(
                f"Failed to add a listener on '{load_balancer_name}' for target
group '{target_group['TargetGroupName']}'."
            )

            if error_code == "ListenerNotFoundException":
                log.error(
                    f"The listener could not be found for the load balancer
'{load_balancer_name}'. "
                    "Please check the load balancer name and target group
configuration."
                )
            if error_code == "InvalidConfigurationRequestException":
                log.error(
                    f"The configuration provided for the listener on load
balancer '{load_balancer_name}' is invalid. "
                    "Please review the provided protocol, port, and target group
settings."
                )
            log.error(f"Full error:\n\t{err}")


    def delete_load_balancer(self, load_balancer_name) -> None:
        """
        Deletes a load balancer.

        :param load_balancer_name: The name of the load balancer to delete.
        """
        try:
            response = self.elb_client.describe_load_balancers(
                Names=[load_balancer_name]
            )
            lb_arn = response["LoadBalancers"][0]["LoadBalancerArn"]
            self.elb_client.delete_load_balancer(LoadBalancerArn=lb_arn)
            log.info("Deleted load balancer %s.", load_balancer_name)
            waiter = self.elb_client.get_waiter("load_balancers_deleted")
            log.info("Waiting for load balancer to be deleted...")
            waiter.wait(Names=[load_balancer_name])
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(
```

```
                f"Couldn't delete load balancer '{load_balancer_name}'. Error
code: {error_code}, Message: {err.response['Error']['Message']}"
            )

            if error_code == "LoadBalancerNotFoundException":
                log.error(
                    f"The load balancer '{load_balancer_name}' does not exist. "
                    "Please check the name and try again."
                )
            log.error(f"Full error:\n\t{err}")


    def get_endpoint(self, load_balancer_name) -> str:
        """
        Gets the HTTP endpoint of the load balancer.

        :return: The endpoint.
        """
        try:
            response = self.elb_client.describe_load_balancers(
                Names=[load_balancer_name]
            )
            return response["LoadBalancers"][0]["DNSName"]
        except ClientError as err:
            log.error(
                f"Couldn't get the endpoint for load balancer
{load_balancer_name}"
            )
            error_code = err.response["Error"]["Code"]
            if error_code == "LoadBalancerNotFoundException":
                log.error(
                    "Verify load balancer name and ensure it exists in the AWS
console."
                )
            log.error(f"Full error:\n\t{err}")

    @staticmethod
    def verify_load_balancer_endpoint(endpoint) -> bool:
        """
        Verify this computer can successfully send a GET request to the load
balancer endpoint.

        :param endpoint: The endpoint to verify.
        :return: True if the GET request is successful, False otherwise.
```

```
        """
        retries = 3
        verified = False
        while not verified and retries > 0:
            try:
                lb_response = requests.get(f"http://{endpoint}")
                log.info(
                    "Got response %s from load balancer endpoint.",
                    lb_response.status_code,
                )
                if lb_response.status_code == 200:
                    verified = True
                else:
                    retries = 0
            except requests.exceptions.ConnectionError:
                log.info(
                    "Got connection error from load balancer endpoint,
retrying..."
                )
                retries -= 1
                time.sleep(10)
        return verified

    def check_target_health(self, target_group_name: str) -> List[Dict[str,
Any]]:
        """
        Checks the health of the instances in the target group.

        :return: The health status of the target group.
        """
        try:
            tg_response = self.elb_client.describe_target_groups(
                Names=[target_group_name]
            )
            health_response = self.elb_client.describe_target_health(
                TargetGroupArn=tg_response["TargetGroups"][0]["TargetGroupArn"]
            )
        except ClientError as err:
            log.error(f"Couldn't check health of {target_group_name} target(s).")
            error_code = err.response["Error"]["Code"]
            if error_code == "LoadBalancerNotFoundException":
                log.error(
                    "Load balancer associated with the target group was not
found. "
```

```
                    "Ensure the load balancer exists, is in the correct AWS
 region, and "
                    "that you have the necessary permissions to access it.",
                )
            elif error_code == "TargetGroupNotFoundException":
                log.error(
                    "Target group was not found. "
                    "Verify the target group name, check that it exists in the
 correct region, "
                    "and ensure it has not been deleted or created in a different
 account.",
                )
            log.error(f"Full error:\n\t{err}")
        else:
            return health_response["TargetHealthDescriptions"]
```

Crie uma classe que use o DynamoDB para simular um serviço de recomendação.

```
class RecommendationService:
    """
    Encapsulates a DynamoDB table to use as a service that recommends books,
 movies,
    and songs.
    """

    def __init__(self, table_name: str, dynamodb_client: boto3.client):
        """
        Initializes the RecommendationService class with the necessary
 parameters.

        :param table_name: The name of the DynamoDB recommendations table.
        :param dynamodb_client: A Boto3 DynamoDB client.
        """
        self.table_name = table_name
        self.dynamodb_client = dynamodb_client

    def create(self) -> Dict[str, Any]:
        """
```

```
        Creates a DynamoDB table to use as a recommendation service. The table
has a
        hash key named 'MediaType' that defines the type of media recommended,
such as
        Book or Movie, and a range key named 'ItemId' that, combined with the
MediaType,
        forms a unique identifier for the recommended item.

        :return: Data about the newly created table.
        :raises RecommendationServiceError: If the table creation fails.
        """
        try:
            response = self.dynamodb_client.create_table(
                TableName=self.table_name,
                AttributeDefinitions=[
                    {"AttributeName": "MediaType", "AttributeType": "S"},
                    {"AttributeName": "ItemId", "AttributeType": "N"},
                ],
                KeySchema=[
                    {"AttributeName": "MediaType", "KeyType": "HASH"},
                    {"AttributeName": "ItemId", "KeyType": "RANGE"},
                ],
                ProvisionedThroughput={"ReadCapacityUnits": 5,
"WriteCapacityUnits": 5},
            )
            log.info("Creating table %s...", self.table_name)
            waiter = self.dynamodb_client.get_waiter("table_exists")
            waiter.wait(TableName=self.table_name)
            log.info("Table %s created.", self.table_name)
        except ClientError as err:
            if err.response["Error"]["Code"] == "ResourceInUseException":
                log.info("Table %s exists, nothing to be done.", self.table_name)
            else:
                raise RecommendationServiceError(
                    self.table_name, f"ClientError when creating table: {err}."
                )
        else:
            return response

    def populate(self, data_file: str) -> None:
        """
        Populates the recommendations table from a JSON file.

        :param data_file: The path to the data file.
```

```
        :raises RecommendationServiceError: If the table population fails.
        """
        try:
            with open(data_file) as data:
                items = json.load(data)
            batch = [{"PutRequest": {"Item": item}} for item in items]
            self.dynamodb_client.batch_write_item(RequestItems={self.table_name:
batch})
            log.info(
                "Populated table %s with items from %s.", self.table_name,
data_file
            )
        except ClientError as err:
            raise RecommendationServiceError(
                self.table_name, f"Couldn't populate table from {data_file}:
{err}"
            )

    def destroy(self) -> None:
        """
        Deletes the recommendations table.

        :raises RecommendationServiceError: If the table deletion fails.
        """
        try:
            self.dynamodb_client.delete_table(TableName=self.table_name)
            log.info("Deleting table %s...", self.table_name)
            waiter = self.dynamodb_client.get_waiter("table_not_exists")
            waiter.wait(TableName=self.table_name)
            log.info("Table %s deleted.", self.table_name)
        except ClientError as err:
            if err.response["Error"]["Code"] == "ResourceNotFoundException":
                log.info("Table %s does not exist, nothing to do.",
self.table_name)
            else:
                raise RecommendationServiceError(
                    self.table_name, f"ClientError when deleting table: {err}."
                )
```

Crie uma classe que envolva as ações do Systems Manager.

```
class ParameterHelper:
    """
    Encapsulates Systems Manager parameters. This example uses these parameters
 to drive
    the demonstration of resilient architecture, such as failure of a dependency
 or
    how the service responds to a health check.
    """

    table: str = "doc-example-resilient-architecture-table"
    failure_response: str = "doc-example-resilient-architecture-failure-response"
    health_check: str = "doc-example-resilient-architecture-health-check"

    def __init__(self, table_name: str, ssm_client: boto3.client):
        """
        Initializes the ParameterHelper class with the necessary parameters.

        :param table_name: The name of the DynamoDB table that is used as a
 recommendation
                           service.
        :param ssm_client: A Boto3 Systems Manager client.
        """
        self.ssm_client = ssm_client
        self.table_name = table_name

    def reset(self) -> None:
        """
        Resets the Systems Manager parameters to starting values for the demo.
        These are the name of the DynamoDB recommendation table, no response when
 a
        dependency fails, and shallow health checks.
        """
        self.put(self.table, self.table_name)
        self.put(self.failure_response, "none")
        self.put(self.health_check, "shallow")

    def put(self, name: str, value: str) -> None:
        """
        Sets the value of a named Systems Manager parameter.

        :param name: The name of the parameter.
        :param value: The new value of the parameter.
        :raises ParameterHelperError: If the parameter value cannot be set.
```

```
        """
        try:
            self.ssm_client.put_parameter(
                Name=name, Value=value, Overwrite=True, Type="String"
            )
            log.info("Setting parameter %s to '%s'.", name, value)
        except ClientError as err:
            error_code = err.response["Error"]["Code"]
            log.error(f"Failed to set parameter {name}.")
            if error_code == "ParameterLimitExceeded":
                log.error(
                    "The parameter limit has been exceeded. "
                    "Consider deleting unused parameters or request a limit
increase."
                )
            elif error_code == "ParameterAlreadyExists":
                log.error(
                    "The parameter already exists and overwrite is set to False.
"
                    "Use Overwrite=True to update the parameter."
                )
            log.error(f"Full error:\n\t{err}")
```

- Para obter detalhes da API, consulte os tópicos a seguir na Referência da API AWS SDK para Python (Boto3).
  - AttachLoadBalancerTargetGroups
  - CreateAutoScalingGroup
  - CreateInstanceProfile
  - CreateLaunchTemplate
  - CreateListener
  - CreateLoadBalancer
  - CreateTargetGroup
  - DeleteAutoScalingGroup
  - DeleteInstanceProfile
  - DeleteLaunchTemplate
  - DeleteLoadBalancer

- DeleteTargetGroup
- DescribeAutoScalingGroups
- DescribeAvailabilityZones
- DescribeIamInstanceProfileAssociations
- DescribeInstances
- DescribeLoadBalancers
- DescribeSubnets
- DescribeTargetGroups
- DescribeTargetHealth
- DescribeVpcs
- RebootInstances
- ReplaceIamInstanceProfileAssociation
- TerminateInstanceInAutoScalingGroup
- UpdateAutoScalingGroup

Para obter uma lista completa dos guias do desenvolvedor do AWS SDK e exemplos de código, consulteUsando esse serviço com um AWS SDK. Este tópico também inclui informações sobre como começar e detalhes sobre versões anteriores do SDK.

Solucionar problemas do Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling fornece erros específicos e descritivos para ajudar a solucionar problemas. Você pode encontrar as mensagens de erro na descrição das ações de escalabilidade.

Tópicos

- Recuperar uma mensagem de erro de ações de escalabilidade
- Desativar atividades de escala
- Recursos adicionais para solução de problemas
- Solucionar problemas do Amazon EC2 Auto Scaling: falhas ao iniciar instâncias do EC2
- Solucionar problemas do Amazon EC2 Auto Scaling: problemas de AMI
- Solucionar problemas do Amazon EC2 Auto Scaling: problemas do balanceador de carga
- Solucionar problemas do Amazon EC2 Auto Scaling: modelos de execução

Recuperar uma mensagem de erro de ações de escalabilidade

Para recuperar uma mensagem de erro da descrição de atividades de escalabilidade, use o comando describe-scaling-activities. Você tem um registro de atividades de escalabilidade que data de 6 semanas atrás. As ações de escalabilidade são ordenadas por hora de início, com as ações de escalabilidade mais recentes listadas primeiro.

> **Note**
>
> As ações de escalabilidade também são exibidas no histórico de atividades no console do Amazon EC2 Auto Scaling, na guia Activity (Atividades) do grupo do Auto Scaling.

Para ver as ações de escalabilidade de um grupo específico do Auto Scaling, use o comando a seguir.

```
aws autoscaling describe-scaling-activities --auto-scaling-group-name my-asg
```

O exemplo a seguir é de uma resposta, em que `StatusCode` contém o status atual da atividade e `StatusMessage` contém a mensagem de erro.

```
{
```

```
    "Activities": [
        {
            "ActivityId": "3b05dbf6-037c-b92f-133f-38275269dc0f",
            "AutoScalingGroupName": "my-asg",
            "Description": "Launching a new EC2 instance: i-003a5b3ffe1e9358e.  Status
 Reason: Instance failed to complete user's Lifecycle Action: Lifecycle Action with
 token e85eb647-4fe0-4909-b341-a6c42d8aba1f was abandoned: Lifecycle Action Completed
 with ABANDON Result",
            "Cause": "At 2021-01-11T00:35:52Z a user request created an
 AutoScalingGroup changing the desired capacity from 0 to 1.  At 2021-01-11T00:35:53Z
 an instance was started in response to a difference between desired and actual
 capacity, increasing the capacity from 0 to 1.",
            "StartTime": "2021-01-11T00:35:55.542Z",
            "EndTime": "2021-01-11T01:06:31Z",
            "StatusCode": "Cancelled",
            "StatusMessage": "Instance failed to complete user's Lifecycle Action:
 Lifecycle Action with token e85eb647-4fe0-4909-b341-a6c42d8aba1f was abandoned:
 Lifecycle Action Completed with ABANDON Result",
            "Progress": 100,
            "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-
west-2b\"...}",
            "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
        },
     ...
    ]
}
```

Para obter uma descrição dos campos na saída, consulte Atividade na Referência da API do Amazon EC2 Auto Scaling.

Para visualizar as ações de dimensionamento para um grupo excluído

Para visualizar as atividades de escalabilidade após a exclusão do grupo Auto Scaling, adicione `--include-deleted-groups` a opção ao comando describe-scaling-activitiesda seguinte maneira.

```
aws autoscaling describe-scaling-activities --auto-scaling-group-name my-asg --include-
deleted-groups
```

O exemplo a seguir é uma resposta com uma ação de escalabilidade para um grupo excluído.

```
{
```

```
    "Activities": [
        {
            "ActivityId": "e1f5de0e-f93e-1417-34ac-092a76fba220",
            "AutoScalingGroupName": "my-asg",
            "Description": "Launching a new EC2 instance.  Status Reason: Your Spot
 request price of 0.001 is lower than the minimum required Spot request fulfillment
 price of 0.0031. Launching EC2 instance failed.",
            "Cause": "At 2021-01-13T20:47:24Z a user request update of AutoScalingGroup
 constraints to min: 1, max: 5, desired: 3 changing the desired capacity from 0 to 3.
  At 2021-01-13T20:47:27Z an instance was started in response to a difference between
 desired and actual capacity, increasing the capacity from 0 to 3.",
            "StartTime": "2021-01-13T20:47:30.094Z",
            "EndTime": "2021-01-13T20:47:30Z",
            "StatusCode": "Failed",
            "StatusMessage": "Your Spot request price of 0.001 is lower than the
 minimum required Spot request fulfillment price of 0.0031. Launching EC2 instance
 failed.",
            "Progress": 100,
            "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-
west-2b\"...}",
            "AutoScalingGroupState": "Deleted",
            "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
        },
     ...
    ]
}
```

Desativar atividades de escala

Você tem as seguintes opções se precisar investigar um problema sem a interferência de políticas de escala ou ações programadas:

- Impeça que todas as políticas de escala dinâmica e ações programadas façam alterações na capacidade desejada do grupo, suspendendo os processos `AlarmNotification` e `ScheduledActions`. Para obter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.
- Desative as políticas de escala dinâmica individuais para que elas não alterem a capacidade desejada do grupo em resposta a alterações na carga. Para obter mais informações, consulte Desabilitar uma política de escalabilidade para um grupo do Auto Scaling.

- Atualize as políticas individuais de escalabilidade de rastreamento de metas para escalar apenas para fora (adicionar capacidade) desativando a parte de expansão das políticas. Esse método evita que a capacidade desejada do grupo diminua, mas permite que seja aumentada quando a carga aumenta. Para obter mais informações, consulte Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling.
- Atualize sua política de escala preditiva para o modo somente de previsão. Enquanto estiver no modo somente de previsão, a escala preditiva continuará a gerar previsões, mas não aumentará automaticamente a capacidade. Para obter mais informações, consulte Desabilitar uma política de escalabilidade para um grupo do Auto Scaling.

Recursos adicionais para solução de problemas

As páginas a seguir apresentam mais informações para solucionar problemas com o Amazon EC2 Auto Scaling.

- Verificar uma ação de escalabilidade para um grupo do Auto Scaling
- Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling
- Verificações de integridade para instâncias em um grupo do Auto Scaling
- Considerações e limitações dos ganchos do ciclo de vida
- Concluir uma ação do ciclo de vida em um grupo do Auto Scaling
- Fornecer conectividade de rede para suas instâncias do Auto Scaling usando a Amazon VPC
- Remover temporiamente instâncias do grupo do Auto Scaling
- Desabilitar uma política de escalabilidade para um grupo do Auto Scaling
- Suspender e retomar os processos do Amazon EC2 Auto Scaling
- Controlar quais instâncias do Auto Scaling serão terminadas durante uma redução de escala na horizontal
- Excluir infraestrutura do Auto Scaling
- Cotas para recursos e grupos do Auto Scaling

Os seguintes AWS recursos também podem ajudar:

- Tópicos AWS do Amazon EC2 Auto Scaling no Centro de conhecimento
- Perguntas sobre o Amazon EC2 Auto AWS Scaling no re:POST

- Publicações do Amazon EC2 Auto Scaling no blog de computação AWS
- Solução de problemas CloudFormation no Guia CloudFormation do usuário

Geralmente, a solução de problemas requer consulta e descoberta iterativas por um especialista ou de uma comunidade de ajudantes. Se você continuar enfrentando problemas depois de tentar as sugestões desta seção, entre em contato AWS Support (no, clique em Support Console de gerenciamento da AWS, Support Center) ou faça uma pergunta no AWS re:POST usando a tag Amazon EC2 Auto Scaling.

Solucionar problemas do Amazon EC2 Auto Scaling: falhas ao iniciar instâncias do EC2

Esta página fornece informações sobre suas instâncias do EC2 que falham ao ativar, as possíveis causas e as etapas que você pode realizar para resolver o problema.

Para recuperar uma mensagem de erro, consulte Recuperar uma mensagem de erro de ações de escalabilidade.

Quando suas instâncias do EC2 falham ao ativar, você pode obter uma ou mais das seguintes mensagens de erro:

Problemas ao iniciar

- A configuração solicitada não é suportada atualmente.
- O grupo de segurança <nome do grupo de segurança> não existe. Falha ao ativar a instância do EC2.
- O par de chaves <par de chaves associado à sua instância do EC2> não existe. Falha ao ativar a instância do EC2.
- O tipo de instância solicitado (<tipo de instância>) não tem suporte na Zona de disponibilidade solicitada (<Zona de disponibilidade da instância>)...
- Seu preço de solicitação spot de 0,015 é inferior ao preço mínimo de atendimento de solicitação spot exigido de 0,0735...
- Nome de dispositivo inválido <nome do dispositivo> / Carregamento do nome de dispositivo inválido. Falha ao ativar a instância do EC2.
- O valor (<nome associado ao dispositivo de armazenamento de instâncias>) do parâmetro virtualName é inválido... Falha ao ativar a instância do EC2.

- Os mapeamentos de dispositivos de blocos do EBS não são compatíveis com o armazenamento de instâncias. AMIs
- Os grupos de posicionamento não podem ser usados com instâncias do tipo '<instance type>'. Falha ao ativar a instância do EC2.
- Cliente. InternalError: Erro do cliente na inicialização.
- No momento, não temos capacidade de <tipo de instância> suficiente para tipo de instância na zona de disponibilidade solicitada. Falha ao ativar a instância do EC2.
- A reserva solicitada não tem capacidade compatível e disponível suficiente para essa solicitação. Falha ao ativar a instância do EC2.
- Sua reserva do bloco de capacidade <reservation id> ainda não está ativa. Falha ao ativar a instância do EC2.
- Não há capacidade spot disponível que corresponda à sua solicitação. Falha ao ativar a instância do EC2.
- <número de instâncias> instância(s) já estão em execução. Falha ao ativar a instância do EC2.

A configuração solicitada não é suportada atualmente.

Causa: algumas opções em seu modelo de execução ou configuração de execução podem não ser compatíveis com o tipo de instância, ou a configuração da instância pode não ser suportada na AWS região ou nas zonas de disponibilidade solicitadas.

Solução: Tente uma configuração de instância diferente. Para pesquisar um tipo de instância que atenda aos seus requisitos, consulte Como encontrar tipos de instâncias do Amazon EC2 no Manual do usuário do Amazon EC2.

Para obter mais orientações sobre como resolver esse problema, verifique o seguinte:

- Certifique-se de que você escolheu uma AMI que é suportada pelo seu tipo de instância. Por exemplo, se o tipo de instância usa um processador AWS Graviton baseado em ARM em vez de um processador Intel Xeon, você precisa de uma AMI compatível com ARM. Para obter mais informações sobre como escolher um tipo de instância compatível, consulte Compatibilidade para alterar o tipo de instância no Guia do usuário do Amazon EC2.
- Teste se o tipo de instância está disponível na região e nas zonas de disponibilidade solicitadas. Os tipos de instância de geração mais recente podem ainda não estar disponíveis em uma determinada região ou zona de disponibilidade. Os tipos de instância da geração anterior podem não estar disponíveis em regiões e zonas de disponibilidade mais recentes. Para pesquisar tipos

de instância oferecidos por local (região ou zona de disponibilidade), use o comando describe-instance-type-offerings. Para obter mais informações, consulte Tipos de nomes de instâncias do Amazon EC2 no Guia do usuário do Amazon EC2.

- Se você usa instâncias dedicadas ou hosts dedicados, verifique se você escolheu um tipo de instância que pode ser usado como uma instância dedicada ou um host dedicado.

O grupo de segurança <nome do grupo de segurança> não existe. Falha ao ativar a instância do EC2.

Causa: o grupo de segurança especificado no modelo de execução ou na configuração de execução pode ter sido excluído.

Solução:

1. Use o comando describe-security-groups para obter a lista dos grupos de segurança associados à sua conta.
2. Na lista, selecione os security groups a serem usados. Para criar um grupo de segurança, use o comando create-security-group.
3. Crie um novo modelo de execução ou uma nova configuração de execução.
4. Atualize seu grupo do Auto Scaling com o novo modelo de execução ou a nova configuração de execução usando o comando update-auto-scaling-group.

O par de chaves <par de chaves associado à sua instância do EC2> não existe. Falha ao ativar a instância do EC2.

Causa: O par de chaves que foi usado ao ativar a instância pode ter sido excluído.

Solução:

1. Use o comando describe-key-pairs para obter a lista de pares de chaves disponíveis para você.
2. Na lista, selecione o par de chaves a ser usado. Para criar um par de chaves, use o comando create-key-pair.
3. Crie um novo modelo de execução ou uma nova configuração de execução.
4. Atualize seu grupo do Auto Scaling com o novo modelo de execução ou a nova configuração de execução usando o comando update-auto-scaling-group.

O tipo de instância solicitado (<tipo de instância>) não tem suporte na Zona de disponibilidade solicitada (<Zona de disponibilidade da instância>)...

Mensagem de erro: O tipo de instância solicitado (<tipo de instância>) não tem suporte na zona de disponibilidade solicitada (<zona de disponibilidade da instância>)…Falha na execução da instância do EC2.

Causa: as zonas de disponibilidade especificadas em seu grupo do Auto Scaling não são compatíveis com o tipo de instância escolhido.

Solução:

1. Verifique quais zonas de disponibilidade oferecem suporte ao tipo de instância escolhido usando o describe-instance-type-offeringscomando ou no console do Amazon EC2 verificando o valor das zonas de disponibilidade no painel de rede da página Tipos de instância.
2. Atualize ou remova a sub-rede de qualquer zona não suportada nas configurações do seu grupo de Auto Scaling usando o comando. update-auto-scaling-group Para obter mais informações, consulte Adicione uma Zona de disponibilidade.

Seu preço de solicitação spot de 0,015 é inferior ao preço mínimo de atendimento de solicitação spot exigido de 0,0735...

Causa: o preço máximo spot na solicitação é inferior ao preço spot do tipo de instância que você selecionou.

Solução: envie uma nova solicitação com um preço máximo spot mais alto (possivelmente o preço sob demanda). Anteriormente, o preço spot pago era baseado em lances. Hoje, você paga o preço spot atual. Ao definir seu preço máximo mais alto, a chance do serviço spot do Amazon EC2 iniciar e manter a quantidade necessária de capacidade é maior.

Nome de dispositivo inválido <nome do dispositivo> / Carregamento do nome de dispositivo inválido. Falha ao ativar a instância do EC2.

Causa 1: os mapeamentos de dispositivos de blocos em seu modelo de execução ou configuração de execução podem conter nomes de dispositivos de blocos que não estão disponíveis ou são incompatíveis no momento.

Solução:

1. Verifique quais nomes de dispositivos estão disponíveis para sua configuração de instância específica. Para mais detalhes sobre nomeação de dispositivos, consulte Device names on Linux instances (Nomenclatura de dispositivos em instâncias do Linux) no Guia do usuário do Amazon EC2.
2. Crie manualmente uma instância do Amazon EC2 que não faça parte do grupo do Auto Scaling e investigue o problema. Se a configuração de nomenclatura do dispositivo de blocos entrar em conflito com os nomes na imagem de máquina da Amazon (AMI), a instância falhará durante o execução. Para mais informações, consulte Mapeamento de dispositivos de blocos no Guia do usuário do Amazon EC2.
3. Depois de confirmar se sua instância foi executada com êxito, use o comando describe-volumes para ver como os volumes estão expostos para a instância.
4. Crie um novo modelo ou uma nova configuração de execução usando o nome do dispositivo listado na descrição do volume.
5. Atualize seu grupo do Auto Scaling com o novo modelo de execução ou a nova configuração de execução usando o comando update-auto-scaling-group.

O valor (<nome associado ao dispositivo de armazenamento de instâncias>) do parâmetro virtualName é inválido... Falha ao ativar a instância do EC2.

Causa: O formato especificado para o nome virtual associado ao dispositivo de blocos está incorreto.

Solução:

1. Crie um novo modelo ou uma nova configuração de execução especificando o nome do dispositivo no parâmetro `virtualName`. Para obter informações sobre o formato dos nomes de dispositivos, consulte Nomenclatura de dispositivos em instâncias do Linux no Manual do usuário do Amazon EC2.
2. Atualize seu grupo do Auto Scaling com o novo modelo de execução ou a nova configuração de execução usando o comando update-auto-scaling-group.

Os mapeamentos de dispositivos de blocos do EBS não são compatíveis com o armazenamento de instâncias. AMIs

Causa: os mapeamentos de dispositivos de blocos especificados no modelo ou na configuração de execução não são compatíveis com sua instância.

Solução:

1. Crie um novo modelo ou uma nova configuração de execução com mapeamentos de dispositivos de blocos suportados pelo seu tipo de instância. Para mais informações, consulte Mapeamento de dispositivo de blocos na Guia de Usuário Amazon EC2.
2. Atualize seu grupo do Auto Scaling com o novo modelo de execução ou a nova configuração de execução usando o comando update-auto-scaling-group.

Os grupos de posicionamento não podem ser usados com instâncias do tipo '<instance type>'. Falha ao ativar a instância do EC2.

Causa: Seu placement group de cluster contém um tipo de instância inválido.

Solução:

1. Para obter mais informações sobre os tipos de instância válidos suportados pelos grupos de colocação, consulte Grupos de colocação no Manual do usuário do Amazon EC2.
2. Siga as instruções detalhadas em Grupos de colocação para criar um novo grupo de colocação.
3. Como alternativa, crie um novo modelo ou uma nova configuração de execução com o tipo de instância suportado.
4. Atualize seu grupo do Auto Scaling com um novo modelo de execução de grupo de colocação ou a nova configuração de execução usando o comando update-auto-scaling-group.

Cliente. InternalError: Erro do cliente na inicialização.

Problema: O Amazon EC2 Auto Scaling tenta iniciar uma instância que tem um volume do EBS criptografado, mas a função vinculada ao serviço não tem acesso à chave gerenciada pelo cliente usada AWS KMS para criptografá-la. Para obter mais informações, consulte Política de AWS KMS chaves necessária para uso com volumes criptografados.

Causa 1: você precisa de uma política de chaves que conceda permissão para usar a chave gerenciada pelo cliente para o perfil vinculado ao serviço adequado.

Solução 1: permita que o perfil vinculado ao serviço use a chave gerenciada pelo cliente da seguinte forma:

1. Determine qual perfil vinculado ao serviço deve ser usado para esse grupo do Auto Scaling.
2. Atualize a política de chaves na chave gerenciada pelo cliente e permita que o perfil vinculado ao serviço use a chave gerenciada pelo cliente.
3. Atualize o grupo do Auto Scaling para usar o perfil vinculado ao serviço.

Para obter um exemplo de uma política de chave que permita que o perfil vinculado ao serviço use a chave gerenciada pelo cliente, consulte Exemplo 1: seções da política de chaves que permitem acesso à chave gerenciada pelo cliente.

Causa 2: Se a chave gerenciada pelo cliente e o grupo Auto Scaling estiverem em AWS contas diferentes, você precisará configurar o acesso entre contas à chave gerenciada pelo cliente para dar permissão para usar a chave gerenciada pelo cliente para a função vinculada ao serviço adequada.

Solução 2: permita que o perfil vinculado ao serviço na conta externa use a chave gerenciada pelo cliente na conta local da seguinte maneira:

1. Atualize a política de chaves na chave gerenciada pelo cliente para permitir que a conta do grupo do Auto Scaling acesse a chave gerenciada pelo cliente.
2. Defina um usuário ou um perfil do IAM na conta do grupo do Auto Scaling que possa criar uma concessão.
3. Determine que perfil vinculado ao serviço deve ser usada para esse grupo do Auto Scaling.
4. Crie uma concessão para a chave gerenciada pelo cliente com o perfil vinculado ao serviço como o principal favorecido.
5. Atualize o grupo do Auto Scaling para usar o perfil vinculado ao serviço.

Para obter mais informações, consulte Exemplo 2: seções da política de chaves que permitem acesso entre contas à chave gerenciada pelo cliente.

Solução 3: Use uma chave gerenciada pelo cliente na mesma conta da AWS que o grupo do Auto Scaling.

1. Copie e criptografe novamente o snapshot com outra chave gerenciada pelo cliente pertencente à mesma conta que o grupo do Auto Scaling.
2. Permita que o perfil vinculado ao serviço use a nova chave gerenciada pelo cliente. Consulte as etapas da Solução 1.

No momento, não temos capacidade de <tipo de instância> suficiente para tipo de instância na zona de disponibilidade solicitada. Falha ao ativar a instância do EC2.

Mensagem de erro: We currently do not have sufficient <instance type> capacity in the Availability Zone you requested (<requested Availability Zone>) (No momento, não temos capacidade do <tipo de instância> suficiente na zona de disponibilidade solicitada (<zona de disponibilidade solicitada>)). O nosso sistema trabalhará no provisionamento de capacidade adicional. No momento, você pode obter capacidade do <tipo de instância> sem especificar uma Zona de disponibilidade em sua solicitação ou escolher a <lista de Zonas de disponibilidade que oferecem suporte ao tipo de instância no momento>. Falha ao ativar a instância do EC2.

Causa: No momento, a combinação do tipo de instância solicitada e da zona de disponibilidade não é compatível.

Solução: Para resolver esse problema, experimente o seguinte:

- Aguarde alguns minutos para que o Amazon EC2 Auto Scaling tenha capacidade para esse tipo de instância em outras zonas de disponibilidade habilitadas.
- Expanda seu grupo do Auto Scaling para zonas de disponibilidade adicionais. Para obter mais informações, consulte Adicione uma Zona de disponibilidade.
- Siga as práticas recomendadas de uso de um conjunto diversificado de tipos de instância para não depender de um tipo de instância específico. Para obter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

A reserva solicitada não tem capacidade compatível e disponível suficiente para essa solicitação. Falha ao ativar a instância do EC2.

Causa 1: você atingiu o limite do número total de instâncias que pode iniciar com uma reserva de capacidade `targeted` sob demanda.

Solução 1: aumente o número de instâncias que você pode iniciar com a reserva de capacidade `targeted` sob demanda ou use um grupo de reservas de capacidade para que qualquer coisa além da capacidade reservada seja executada como capacidade sob demanda regular. Para obter mais informações, consulte Reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas.

Causa 2: você atingiu o limite do número total de instâncias que pode iniciar com um bloco de capacidade.

Com os Blocos de Capacidade, você fica limitado pela quantidade de capacidade adquirida originalmente. Se você observar um número de inicializações superior ao previsto e usar toda a capacidade disponível, isso causará falha nas inicializações. As instâncias encerradas passam por um longo processo de limpeza antes de serem totalmente encerradas. Durante esse período, elas não podem ser reutilizadas. Isso também pode causar falha nas inicializações. Para obter mais informações, consulte Use Capacity Blocks para workloads de machine learning..

Solução 2: Para resolver esse problema, experimente o seguinte:

- Mantenha a solicitação como está. Se uma instância do tipo Capacity Block estiver sendo finalizada, é necessário aguardar alguns minutos para que a instância termine de ser finalizada e a capacidade fique disponível novamente. O Amazon EC2 Auto Scaling continuará a fazer a solicitação de execução automaticamente até que a capacidade seja disponibilizada.
- Certifique-se de adquirir capacidade suficiente para acomodar sua workload de pico, para que você não encontre esse erro com frequência.

Sua reserva do bloco de capacidade <reservation id> ainda não está ativa. Falha ao ativar a instância do EC2.

Causa: O bloco de capacidade especificado ainda não está ativo.

Solução: siga a abordagem recomendada para blocos de capacidade e use a escalabilidade programada. Isso ajuda a garantir que você aumente a capacidade desejada do grupo do Auto Scaling somente quando a reserva estiver ativa e a diminua antes que a reserva termine.

Não há capacidade spot disponível que corresponda à sua solicitação. Falha ao ativar a instância do EC2.

Causa: no momento, não há capacidade de reserva suficiente para atender à sua solicitação de instâncias spot.

Solução: Para resolver esse problema, experimente o seguinte:

- Aguarde alguns minutos; a capacidade pode mudar com frequência. O Amazon EC2 Auto Scaling continuará a fazer a solicitação de execução automaticamente até que a capacidade seja disponibilizada.
- Expanda seu grupo do Auto Scaling para zonas de disponibilidade adicionais. Para obter mais informações, consulte Adicione uma Zona de disponibilidade.
- Siga as práticas recomendadas de uso de um conjunto diversificado de tipos de instância para não depender de um tipo de instância específico. Para obter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

<número de instâncias> instância(s) já estão em execução. Falha ao ativar a instância do EC2.

Causa: você atingiu o limite do número total de instâncias que pode iniciar em uma região. Quando você cria sua AWS conta, definimos limites padrão para o número de instâncias que você pode executar por região.

Solução: Para resolver esse problema, experimente o seguinte:

- Se os limites atuais não forem adequados às suas necessidades, você poderá solicitar um aumento de cota por região. Para obter mais informações, consulte Amazon EC2 service quotas no Guia do usuário do Amazon EC2.
- Envie uma nova solicitação com um número de instâncias reduzido (que pode ser aumentado posteriormente).

Solucionar problemas do Amazon EC2 Auto Scaling: problemas de AMI

Esta página fornece informações sobre os problemas associados aos seus AMIs, as possíveis causas e as etapas que você pode seguir para resolvê-los.

Para recuperar uma mensagem de erro, consulte Recuperar uma mensagem de erro de ações de escalabilidade.

Quando suas instâncias do EC2 não ativam devido a problemas com sua AMI, você pode obter uma ou mais das seguintes mensagens de erro.

Problemas de AMI

- O ID da AMI <ID de sua AMI> não existe. Falha ao ativar a instância do EC2.
- A AMI <ID da AMI> está pendente e não pode ser executada. Falha ao ativar a instância do EC2.
- Nome do dispositivo inválido <device name>. Falha ao ativar a instância do EC2.
- A arquitetura 'arm64' do tipo de instância especificado não corresponde à arquitetura 'x86_64' da AMI especificada... Falha na execução da instância do EC2.
- A AMI '<AMI ID>' está desabilitada e não pode ser executada. Falha ao ativar a instância do EC2.

> **Important**
>
> AWS suporta o compartilhamento privado de uma AMI com outra AWS conta modificando as permissões da AMI. Se uma AMI se tornar privada sem ser compartilhada, isso pode resultar em um erro de autorização ao iniciar novas instâncias. Para obter mais informações sobre compartilhamento privado AMIs, consulte Compartilhar uma AMI com AWS contas específicas no Guia do usuário do Amazon EC2.

O ID da AMI <ID de sua AMI> não existe. Falha ao ativar a instância do EC2.

- Causa: a AMI pode ter sido excluída depois da criação do modelo de execução ou da configuração de execução.
- Solução:

1. Crie um novo modelo de execução ou uma nova configuração de execução usando uma AMI válida.
2. Atualize seu grupo do Auto Scaling com o novo modelo de execução ou a nova configuração de execução usando o comando update-auto-scaling-group.

A AMI <ID da AMI> está pendente e não pode ser executada. Falha ao ativar a instância do EC2.

Causa: Você pode ter acabado de criar a AMI (usando um snapshot de uma instância em execução ou de qualquer outra maneira) e ela pode não estar disponível ainda.

Solução: é necessário aguardar até que sua AMI esteja disponível e, em seguida, criar um modelo de execução ou uma configuração de execução.

Nome do dispositivo inválido <device name>. Falha ao ativar a instância do EC2.

Causa: Ao conectar um volume do EBS a uma instância do EC2, é necessário fornecer um nome de dispositivo válido para o volume. A AMI selecionada deve ser compatível com esse nome de dispositivo.

Solução:

1. Crie um novo modelo de inicialização ou configuração de inicialização e especifique o nome de dispositivo correto para sua AMI. A convenção de nomenclatura recomendada varia de acordo com o tipo de virtualização da AMI. Para obter mais informações, consulte Nomes de dispositivos no Guia do usuário do Amazon EC2.
2. Atualize seu grupo do Auto Scaling com o novo modelo de execução ou a nova configuração de execução usando o comando update-auto-scaling-group.

A arquitetura 'arm64' do tipo de instância especificado não corresponde à arquitetura 'x86_64' da AMI especificada... Falha na execução da instância do EC2.

Causa 1: Se a arquitetura da AMI e o tipo de instância usado em seu modelo de execução ou configuração de execução não forem os mesmos, você receberá um erro quando o Amazon EC2 Auto Scaling tentar iniciar uma instância usando a configuração de instância incompatível.

Solução 1:

1. Verifique a arquitetura da sua AMI usando o comando describe-images ou do console do Amazon EC2 verificando o valor da arquitetura no painel de detalhes da página Amazon Machine Images (). AMIs
2. Encontre um tipo de instância que tenha a mesma arquitetura da sua AMI usando o describe-instance-typescomando ou no console do Amazon EC2 verificando a coluna Arquitetura na tela Tipos de instância. Para obter mais informações sobre como escolher um tipo de instância compatível, consulte Compatibilidade para alterar o tipo de instância no Guia do usuário do Amazon EC2.
3. Crie um novo modelo ou uma nova configuração de execução usando um tipo de instância que tenha a mesma arquitetura da sua AMI.
4. Atualize seu grupo do Auto Scaling com o novo modelo de execução ou a nova configuração de execução usando o comando update-auto-scaling-group.

Causa 2: O Amazon EC2 Auto Scaling tenta iniciar um tipo de instância especificado na política de instâncias mistas do seu grupo do Auto Scaling, mas o tipo de instância não tem a mesma arquitetura da AMI especificada em seu modelo de execução.

Solução 1: não inclua tipos de instância que tenham arquiteturas diferentes em sua política de instâncias mistas.

1. Verifique a arquitetura da sua AMI usando o comando describe-images ou do console do Amazon EC2 verificando o valor da arquitetura no painel de detalhes da página Amazon Machine Images (). AMIs
2. Verifique a arquitetura de cada tipo de instância que você pretende incluir em sua política de instâncias mistas usando o describe-instance-typescomando ou a partir do console do Amazon EC2, verificando a coluna Arquitetura na tela Tipos de instância. Para obter informações sobre

como escolher um tipo de instância compatível, consulte Compatibilidade para alterar o tipo de instância no Guia do usuário do Amazon EC2.

3. Atualize ou remova os tipos de instância incompatíveis do seu grupo de Auto Scaling usando o update-auto-scaling-groupcomando.

Solução 2: para iniciar instâncias Arm (Graviton2) e x86_64 (Intel) no mesmo grupo do Auto Scaling, é necessário usar modelos de execução que ofereçam suporte a uma AMI compatível com ARM e uma AMI compatível com Intel x86, respectivamente, para corresponder aos tipos de instância em sua política de instâncias mistas.

1. Verifique a arquitetura da AMI em seu modelo de lançamento existente usando o comando describe-images ou no console do Amazon EC2 verificando o valor da arquitetura no painel de detalhes da página Amazon Machine Images (). AMIs
2. Crie um novo modelo de execução usando uma AMI que corresponda à outra arquitetura que você pretende usar.
3. Atualize seu grupo de Auto Scaling para substituir o modelo de execução existente e especificar o novo modelo de execução para cada tipo de instância compatível usando o comando. update-auto-scaling-group Para obter mais informações, consulte Usar vários modelos de execução.

A AMI '<AMI ID>' está desabilitada e não pode ser executada. Falha ao ativar a instância do EC2.

Causa: você está tentando iniciar instâncias de uma AMI que foi desabilitada. Para obter mais informações, consulte Desabilitar uma AMI no Guia do usuário do Amazon EC2.

Solução:

1. Crie um novo modelo ou uma nova configuração de execução e especifique uma AMI que não esteja desabilitada.
2. Atualize seu grupo do Auto Scaling com o novo modelo de execução ou a nova configuração de execução usando o comando update-auto-scaling-group.

Solucionar problemas do Amazon EC2 Auto Scaling: problemas do balanceador de carga

Esta página fornece informações sobre os problemas causados pelo balanceador de carga associados a seu grupo do Auto Scaling, as possíveis causas e as etapas que você pode realizar para resolver os problemas.

Para recuperar uma mensagem de erro, consulte Recuperar uma mensagem de erro de ações de escalabilidade.

Quando houver falha ao iniciar suas instâncias do EC2 devido a problemas com o balanceador de carga associados a seu grupo do Auto Scaling, você poderá receber uma ou mais das seguintes mensagens de erro.

Problemas do balanceador de carga

- Um ou mais grupos de destino não encontrados. Falha na validação da configuração do balanceador de carga.
- Não é possível encontrar o Load Balancer <seu load balancer>. Falha na validação da configuração do balanceador de carga.
- Não há nenhum balanceador de carga ATIVO chamado <nome do balanceador de carga>. Falha ao atualizar a configuração do balanceador de carga.
- A instância do EC2 <ID da instância> não está na VPC. Falha ao atualizar a configuração do balanceador de carga.

> **Note**
>
> Você pode usar o Reachability Analyzer para solucionar problemas de conectividade verificando se as instâncias do seu grupo do Auto Scaling podem ser acessadas por meio do balanceador de carga. Para saber mais sobre os diferentes problemas de configuração incorreta de rede que são automaticamente detectados pelo Reachability Analyzer, consulte Reachability Analyzer no Guia do usuário do Reachability Analyzer.

Um ou mais grupos de destino não encontrados. Falha na validação da configuração do balanceador de carga.

Problema: quando seu grupo do Auto Scaling inicia instâncias, o Amazon EC2 Auto Scaling tenta validar a existência dos recursos do Elastic Load Balancing associados ao grupo do Auto Scaling. Quando um grupo-alvo não pode ser encontrado, a atividade de escalabilidade falha e você obtém o erro `One or more target groups not found. Validating load balancer configuration failed.`.

Causa 1: um grupo-alvo vinculado ao seu grupo do Auto Scaling foi excluído.

Solução 1: Você pode criar um novo grupo de Auto Scaling sem o grupo-alvo ou remover o grupo-alvo não utilizado do grupo Auto Scaling usando o console do Amazon EC2 Auto Scaling ou o comando -groups. detach-load-balancer-target

Causa 2: O grupo-alvo existe, mas houve um problema ao tentar especificar o ARN do grupo-alvo ao criar o grupo doAuto Scaling. Os recursos não estão criados na ordem correta.

Solução 2: crie um novo grupo do Auto Scaling e especifique o nome do balanceador de carga no final.

Não é possível encontrar o Load Balancer <seu load balancer>. Falha na validação da configuração do balanceador de carga.

Problema: quando seu grupo do Auto Scaling inicia instâncias, o Amazon EC2 Auto Scaling tenta validar a existência dos recursos do Elastic Load Balancing associados ao grupo do Auto Scaling. Quando não é possível encontrar um Classic Load Balancer, a atividade de escalabilidade falha e você recebe o erro `Cannot find Load Balancer <your load balancer>. Validating load balancer configuration failed.`.

Causa 1: o Classic Load Balancer foi excluído.

Solução 1: Você pode criar um novo grupo de Auto Scaling sem o balanceador de carga ou remover o balanceador de carga não utilizado do grupo Auto Scaling usando o console do Amazon EC2 Auto Scaling ou o comando. detach-load-balancers

Causa 2: o Classic Load Balancer existe, mas houve um problema ao tentar especificar o nome do balanceador de carga durante a criação do grupo do Auto Scaling. Os recursos não estão criados na ordem correta.

Solução 2: crie um novo grupo do Auto Scaling e especifique o nome do balanceador de carga no final.

Não há nenhum balanceador de carga ATIVO chamado <nome do balanceador de carga>. Falha ao atualizar a configuração do balanceador de carga.

Causa: O balanceador de carga especificado pode ter sido excluído.

Solução: você pode criar um novo balanceador de carga e, em seguida, criar um novo grupo do Auto Scaling ou criar um novo grupo do Auto Scaling sem o balanceador de carga.

A instância do EC2 <ID da instância> não está na VPC. Falha ao atualizar a configuração do balanceador de carga.

Causa: A instância especificada não existe na VPC.

Solution: você pode excluir o balanceador de carga associado à instância ou criar um novo grupo do Auto Scaling.

Solucionar problemas do Amazon EC2 Auto Scaling: modelos de execução

Use as informações a seguir para ajudar a diagnosticar e corrigir problemas comuns que podem ser encontrados ao tentar especificar um modelo de inicialização para o grupo do Auto Scaling.

Não é possível iniciar instâncias

Se você não conseguir iniciar instâncias com um modelo de inicialização já especificado, verifique o seguinte para a solução de problemas em geral: Solucionar problemas do Amazon EC2 Auto Scaling: falhas ao iniciar instâncias do EC2.

É necessário usar um modelo de inicialização totalmente formado válido (valor inválido)

Problema: quando você tenta especificar um modelo de inicialização para um grupo do Auto Scaling, recebe o erro `You must use a valid fully-formed launch template` (Você não está autorizado a usar o modelo de inicialização). Você pode encontrar esse erro porque os valores

no modelo de inicialização só são validados quando um grupo do Auto Scaling que está usando o modelo de inicialização é criado ou atualizado.

Causa 1: se você receber um erro `You must use a valid fully-formed launch template` (É necessário usar um modelo de inicialização totalmente formado válido), existem problemas que fazem com que o Amazon EC2 Auto Scaling considere inválido algum detalhe do modelo de inicialização. Esse é um erro genérico que pode ter várias causas diferentes.

Solução 1: tente as seguintes etapas para solucionar os problemas:

1. Preste atenção na segunda parte da mensagem de erro para encontrar mais informações. Após o erro `You must use a valid fully-formed launch template` (É necessário usar um modelo de inicialização totalmente formado válido), veja a mensagem de erro mais específica que identifica o problema que você precisa resolver.
2. Se você não conseguir encontrar a causa, teste o modelo de execução com o comando run-instances. Use a opção `--dry-run`, como mostrado no exemplo a seguir. Isso permite reproduzir o problema e pode fornecer insights sobre a causa do mesmo.

   ```
   aws ec2 run-instances --launch-template LaunchTemplateName=my-template,Version='1' --
   dry-run
   ```

3. Se um valor não for válido, verifique se o recurso especificado existe e se está correto. Por exemplo, quando você especificar um par de chaves do Amazon EC2, o recurso deverá existir na conta e na região em que você estiver criando ou atualizando o grupo do Auto Scaling.
4. Se as informações esperadas estiverem ausentes, verifique as configurações e ajuste o modelo de inicialização conforme necessário.
5. Depois de fazer as alterações, execute novamente o comando run-instances com a opção `--dry-run` para verificar se o modelo de execução usa valores válidos.

Para obter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.

Você não está autorizado a usar o modelo de execução (permissões insuficientes)

Problema: quando você tenta especificar um modelo de inicialização para um grupo do Auto Scaling, recebe o erro `You are not authorized to use launch template` (Você não está autorizado a usar o modelo de inicialização).

Causa 1: se você estiver tentando usar um modelo de execução e as credenciais do IAM não tiverem permissões suficientes, você receberá um erro informando que não está autorizado a usar o modelo de execução.

Solução 1: Para resolver esse problema, experimente o seguinte:

- Verifique se as credenciais do IAM que você está usando para fazer a solicitação têm permissões para chamar as ações de API do EC2 necessárias, incluindo a ação`ec2:RunInstances`. Se você especificou qualquer tag no seu modelo de execução, também deverá ter permissão para usar a ação`ec2:CreateTags`.
- Como alternativa, verifique se as credenciais do IAM que você está usando para fazer a solicitação estão atribuídas à política`AmazonEC2FullAccess`. Essa política AWS gerenciada concede acesso total a todos os recursos do Amazon EC2 e serviços relacionados, incluindo Amazon EC2 Auto Scaling e CloudWatch Elastic Load Balancing.

Para obter mais informações sobre as permissões necessárias para usar modelos de execução, incluindo exemplos de políticas do IAM, consulte Controlar o acesso aos modelos de execução com permissões do IAM no Guia do usuário do Amazon EC2. Para obter exemplos de políticas do IAM, consulte Controle o uso do modelo de execução do Amazon EC2 em grupos do Auto Scaling.

Causa 2: se estiver tentando usar um modelo de execução que especifica um perfil da instância, será necessário ter permissão do IAM para transmitir o perfil do IAM que está associado ao perfil da instância.

Solução 2: verifique se as credenciais do IAM que você está usando para fazer a solicitação têm a `iam:PassRole` permissão correta para transmitir o perfil especificado ao serviço Amazon EC2 Auto Scaling. Para obter mais informações e um exemplo de política do IAM, consulte Perfis do IAM para aplicações que são executadas em instâncias do Amazon EC2. Para obter mais tópicos de solução de problemas relacionados a perfis de instância, consulte Solução de problemas do Amazon EC2 e IAM no Manual do usuário do IAM.

Causa 3: Se você estiver tentando usar um modelo de execução que especifica uma AMI em outra Conta da AWS, e a AMI é privada e não é compartilhada com a pessoa que Conta da AWS você está usando, você recebe uma mensagem de erro informando que não está autorizado a usar o modelo de execução.

Solução 3: verifique se as permissões na AMI incluem a conta que você está usando. Para obter mais informações, consulte Share an AMI with specific Contas da AWS (Compartilhar uma AMI com específico) no Guia do usuário do Amazon EC2.

Informações relacionadas

Os recursos relacionados a seguir podem ajudar você à medida que trabalha com este serviço.

Recurso	Description
Referência da API do Amazon EC2 Auto Scaling	A documentação de cada operação de API mostra os parâmetros de solicitação e a resposta XML e fornece links para tópicos de referência do SDK específicos da linguagem.
escalonamento automático na AWS CLI referência de comando	Descrições dos AWS CLI comandos que você pode usar para trabalhar com grupos do Auto Scaling.
Ferramentas da AWS para PowerShell Referência de cmdlet	As AWS Ferramentas PowerShell permitem que você crie scripts de operações em seus AWS recursos a partir da linha de PowerShell comando.
Crie grupos de Auto Scaling com AWS CloudFormation	O recurso AWS::AutoScaling::AutoScalingAWS::AutoScaling::AutoScalingGroup permite criar, modelar e gerenciar seus grupos de Auto Scaling sem ações manuais.
Endpoints e cotas do Amazon EC2 Auto Scaling no Referência geral da AWS	Informações sobre regiões e endpoints do Amazon EC2 Auto Scaling.
Páginas do produtos	A principal página da Web para obter informaçõ es sobre o Amazon EC2 Auto Scaling.
AWS re: Post	AWS serviço gerenciado de perguntas e respostas (Q&A) que oferece respostas de crowdsourcing e revisadas por especialistas para suas perguntas técnicas.

Recurso	Description
Crie uma AMI no Guia do usuário do Amazon EC2	Aprenda como criar uma imagem de máquina da Amazon (AMI) a partir de uma instância personalizada.
Conexão à sua instância de Linux no Guia do usuário do Amazon EC2	Saiba como se conectar às instâncias do Linux que você executa.
Conexão à sua instância do Windows no Guia do usuário do Amazon EC2	Saiba como se conectar às instâncias do Windows que você inicia.
Criação de um alarme de cobrança para monitorar suas AWS cobranças estimadas no Guia do CloudWatch usuário da Amazon	Saiba como monitorar suas cobranças estimadas usando CloudWatch.
Guia do usuário do Application Auto Scaling	Saiba como configurar a escalabilidade automática para recursos escalavéis além do Amazon EC2.

Histórico do documento

A tabela a seguir descreve adições importantes feitas na documentação do Amazon EC2 Auto Scaling, a partir de julho de 2018. Para receber notificações sobre atualizações dessa documentação, você pode se inscrever em o feed RSS.

Alteração	Descrição	Data
Atualização de segurança do IAM	AutoScalingServiceRolePolicyAgora inclui a `ec2:CreateReplaceRootVolumeTask` permissão que permite substituir volumes raiz em instâncias como parte de uma atualização de instância. Além disso, a `events:DescribeRule` permissão foi removida da política.	20 de novembro de 2025
Substitua os volumes raiz	A substituição do volume raiz atualiza suas instâncias substituindo somente o volume raiz do EBS enquanto mantém a instância em execução. Para obter mais informações, consulte Substituir volumes raiz durante a atualização da instância.	20 de novembro de 2025
Políticas de ciclo de vida da instância	As políticas de ciclo de vida da instância permitem que você controle o que acontece quando os ganchos do ciclo de vida de encerramento não são concluídos com êxito. Para obter mais informações, consulte Substituir volumes	20 de novembro de 2025

	raiz durante a atualização da instância.	
Provisionamento síncrono	Agora você pode usar a LaunchInstances API para iniciar instâncias de forma síncrona com feedback imediato e ter controle preciso sobre em quais instâncias da Zona de Disponibilidade são iniciadas. Para obter mais informações, consulte Provisionamento síncrono.	18 de novembro de 2025
Destino de notificações do Lambda para ganchos do ciclo de vida	Agora você pode encaminhar as notificações de ganchos do ciclo de vida diretamente para AWS Lambda as funções. Para obter mais informações, consulte Encaminhar notificações AWS Lambda diretamente para.	28 de julho de 2025
Métricas de alta resolução	O rastreamento de metas agora oferece suporte a CloudWatch métricas de alta resolução com pontos de dados de segundo nível que são publicados em intervalos menores que um minuto. Para obter mais informações, consulte Criação de uma política de rastreamento de destino usando métricas de alta resolução para uma resposta mais rápida.	22 de novembro de 2024

Atualização de segurança do IAM	A política AutoScalingServiceRolePolicygerenciada agora concede permissão adicional aos Resource Groups`resource-groups:ListGroupResources` .	20 de novembro de 2024
Proteção de performance	Ao usar a seleção de tipo de instância baseada em atributos para o seu grupo do Auto Scaling, agora é possível ativar a proteção de performance para garantir que os tipos de instância selecionados sejam semelhantes ou excedam uma linha de base de performance especificada. Para obter mais informações, consulte Criação de grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos.	20 de novembro de 2024
Preferência de reserva de capacidade	Agora é possível priorizar a execução de instâncias em reservas de capacidade. Para obter mais informações, consulte Reservas de capacidade.	20 de novembro de 2024

Mudança de zona	Agora é possível usar a mudança de zona para a recuperação de problemas de aplicações em uma zona de disponibilidade. Para obter mais informações, consulte Mudança de zona de Grupo do Auto Scaling.	18 de novembro de 2024
Distribuição de zonas de disponibilidade	Agora é possível escolher uma distribuição de zona de disponibilidade para seu grupo do Auto Scaling. Para obter mais informações, consulte Distribuição de zonas de disponibilidade de grupo do Auto Scaling.	7 de novembro de 2024
Atualização de segurança do IAM	A política AutoScalingServiceRolePolicygerenciada agora concede permissões adicionais ao Amazon EC2 (`ec2:GetSecurityGroupsForVpc` `eec2:GetInstanceTypesFromInstanceRequirements`).	29 de fevereiro de 2024

Hibernação de piscina quente suportada em adicionais Regiões da AWS	Agora você pode hibernar instâncias em um pool aquecido em duas regiões adicionais: AWS GovCloud (Leste dos EUA) e AWS GovCloud (Oeste dos EUA). Para obter mais informações sobre pools ativos, consulte Pools ativos para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	26 de fevereiro de 2024
Hibernação de piscina quente suportada em adicionais Regiões da AWS	Agora você pode hibernar instâncias em um pool ativo em mais duas regiões: Europa (Zurique) e Oriente Médio (Emirados Árabes Unidos). Para obter mais informações sobre pools ativos, consulte Pools ativos para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	21 de fevereiro de 2024
Suporte para o uso de parâmetros entre contas	Agora você pode usar um AWS Systems Manager parâmetro compartilhado de outro Conta da AWS com o Amazon EC2 Auto Scaling. Para obter mais informações, consulte Usar AWS Systems Manager parâmetros em vez de AMI IDs em modelos de execução no Guia do usuário do Amazon EC2 Auto Scaling.	21 de fevereiro de 2024

Nova opção de proteção de preço spot	Agora você pode definir o limite de proteção de preço para instâncias spot como uma porcentagem de um preço sob demanda ao usar a seleção de tipo de instância baseada em atributos. Para obter mais informações, consulte Proteção de instâncias na Guia do usuário do Amazon EC2 Auto Scaling.	29 de janeiro de 2024
Políticas de manutenção de instância	Agora você pode usar uma política de manutenção de instâncias para definir se as instâncias são executadas antes ou depois do encerramento das instâncias existentes durante eventos que fazem com que suas instâncias sejam substituídas, incluindo uma atualização da instância. Para obter mais informações, consulte Políticas de manutenção de instância no Guia do usuário do Amazon EC2 Auto Scaling.	15 de novembro de 2023

Blocos de capacidade para ML	Agora você pode iniciar instâncias em um bloco de capacidade especificando o ID de reserva do bloco de capacidade ao criar um modelo de execução. Com os blocos de capacidade, você pode reservar instâncias de GPU em uma data futura para apoiar suas cargas de trabalho de machine learning de curta duração (ML). Para obter mais informações, consulte Usar blocos de capacidade para workloads de machine learning no Guia do usuário do Amazon EC2 Auto Scaling.	31 de outubro de 2023
Novos recursos de atualização de instância	Agora você pode configurar uma atualização de instância para definir seu status como falha e, opcionalmente, reverter quando detectar que um CloudWatch alarme especificado entrou no estado. `ALARM` Para obter mais informações, consulte Desfazer alterações com uma reversão no Guia do usuário do Amazon EC2 Auto Scaling.	31 de julho de 2023

Alterações do guia	Um novo tópico sobre a execução de instâncias sob demanda nas reservas de capacidade foi adicionad o ao guia. Para obter mais informações, consulte Use reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas no guia do usuário de Amazon EC2 Auto Scaling.	28 de julho de 2023
Alterações do guia	Um novo tópico sobre como migrar suas AWS CloudForm ation pilhas das configura ções de lançamento para os modelos de lançamento foi adicionado ao guia. Para obter mais informações, consulte Migre AWS CloudFormation Stacks das configurações de lançamento para os modelos de lançamento no Guia do usuário do Amazon EC2 Auto Scaling.	18 de abril de 2023

Suporte para novas operações de API	Esta versão adiciona três novas operações de API,,`AttachTrafficSources` `DetachTrafficSources` e `DescribeTrafficSources` . Além disso, um novo campo `TrafficSources` foi adicionado aos resultados das `DescribeAutoScalingGroups` operações. Um novo status de atividade ,,`WaitingForConnectionDraining` foi adicionado aos resultados das operações `DescribeScalingActivities` . O Amazon EC2 Auto Scaling também oferece suporte a um novo valor,,`VPC_LATTICE` para o campo `HealthCheckType` nas operações,`CreateAutoScalingGroup` `UpdateAutoScalingGroup` e `DescribeAutoScalingGroups` . Para obter mais informações, consulte a Referência da API do Amazon EC2 Auto Scaling.	31 de março de 2023

Suporte para o Amazon VPC Lattice	É a versão de disponibilidade geral do VPC Lattice para o Amazon EC2 Auto Scaling. Para obter mais informaçõ es, consulte o tráfego de rota para o seu grupo de escala automático com um grupo de destino de treliça VPC no guia do usuário do Amazon EC2 Auto Scaling.	31 de março de 2023
Alterações do guia	A seção com AWS CLI exemplos para trabalhar com o Elastic Load Balancing agora inclui exemplos novos e atualizados. Para obter mais informações, consulte Exemplos de como trabalhar com o Elastic Load Balancing com o AWS Command Line Interface (AWS CLI) no Guia do usuário do Amazon EC2 Auto Scaling.	31 de março de 2023
Support para escalabilidade preditiva, além de Regiões da AWS	Agora você pode criar políticas de escalabilidade preditiva nas regiões do Oriente Médio (EAU) e AWS GovCloud (Leste dos EUA). Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	16 de março de 2023

Novos recursos de atualização de instância	Agora você pode optar por terminar ou ignorar instâncias em espera e substituir ou ignorar instâncias protegidas de redução da escala horizontalmente, em vez de esperar que elas se tornem substituíveis. Também é possível reverter as alterações de uma atualização de instância com falha. Como parte da atualização, a documentação foi expandida para incluir tópicos sobre reverter uma atualização de instância, cancelar uma atualização de instância e entender os valores padrão dos parâmetros configuráveis de uma atualização de instância. Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na atualização de uma instância no Manual do usuário do Amazon EC2 Auto Scaling.	10 de fevereiro de 2023

Support para usar um AWS Systems Manager parâmetro para uma ID de AMI	Agora você pode usar um parâmetro do Systems Manager em vez de um ID de AMI no modelo de execução. Para obter mais informaçõ es, consulte Uso de AWS Systems Manager parâmetro s em vez de AMI IDs em modelos de execução no Guia do usuário do Amazon EC2 Auto Scaling.	19 de janeiro de 2023
Recomendações de escalabil idade preditiva	Já é possível obter recomenda ções para avaliar e escolher a política de escalabilidade preditiva correta no console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte Avaliar políticas de escalabilidade preditiva no Guia do usuário do Amazon EC2 Auto Scaling.	18 de janeiro de 2023
Previsões de escalabilidade preditiva	As previsões geradas pelo dimensionamento preditivo agora são atualizadas a cada seis horas, em vez de diariamente. Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	6 de janeiro de 2023

Support para matemática CloudWatch métrica	Agora você pode usar a matemática métrica ao criar políticas de dimensionamento de rastreamento de destino. Com a matemática métrica, você pode consultar várias CloudWatch métricas e usar expressões matemáticas para criar novas séries temporais com base nessas métricas. Para obter mais informações, consulte Escalabilidade para o Amazon EC2 Auto Scaling usando a escalabilidade para o Amazon EC2 Auto Scaling usando matemática de escalabilidade para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	8 de dezembro de 2022
Atualizar permissões de perfil vinculado a serviços do IAM	A política `AutoScalingServiceRolePolicy` concede permissões adicionais ao Amazon EC2 Auto Scaling. Para obter mais informações, consulte Políticas gerenciadas pela AWS para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	6 de dezembro de 2022

Nova estratégia de alocação spot	Agora você pode usar a estratégia de alocação otimizada de preço e capacidade para solicitar instâncias spot dos pools spot com menor probabilidade de interrupção e com o preço mais baixo possível. Para obter mais informações, consulte Allocation strategies (Estratégias de alocação) no Guia do usuário do Amazon EC2 Auto Scaling.	10 de novembro de 2022
Compatibilidade com escalação preditiva na região Ásia-Pacífico (Jacarta)	Agora você pode criar políticas de escalação preditiva na Região Ásia-Pacífico (Jacarta) . Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	13 de outubro de 2022
Compatibilidade com métricas personalizadas para escalação preditiva no console	Agora você pode usar métricas personalizadas ao criar políticas de escalação preditiva no console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	13 de outubro de 2022

CloudWatch monitoramento de métricas de escalabilidade preditiva	Agora você pode acessar dados de monitoramento para escalabilidade preditiva usando. CloudWatch Isso permite que você use a matemática métrica para criar novas séries temporais que exibem a precisão dos dados de previsão. Para obter mais informações, consulte Monitore métricas de escalabilidade preditiva CloudWatch no Guia do usuário do Amazon EC2 Auto Scaling.	7 de julho de 2022
Compatibilidade com escalação preditiva na região Ásia-Pacífico (Osaka)	Agora você pode criar políticas de escalação preditiva na Região Ásia-Pacífico (Osaka). Para obter mais informaçõ es, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	6 de julho de 2022
Compatibilidade com hibernação de pool ativo passa a ser oferecida em regiões adicionais	Agora você pode hibernar instâncias em um pool ativo em mais quatro regiões: África (Cidade do Cabo), Ásia-Pací fico (Jacarta), Ásia-Pacífico (Osaka) e Europa (Milão). Para obter mais informações sobre pools ativos, consulte Pools ativos para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	5 de julho de 2022

Atualização das verificações de integridade	Ao executar verificações de integridade, agora o Amazon EC2 Auto Scaling ajuda a minimizar qualquer tempo de inatividade que possa ocorrer devido a problemas temporários ou verificações de integridade mal configuradas. Para obter mais informaçõ es, consulte Como o Amazon EC2 Auto Scaling minimiza o tempo de inatividade no Guia do usuário do Amazon EC2 Auto Scaling.	21 de maio de 2022
Aquecimento de instância padrão	Agora é possível unificar todas as configurações de aquecimento e desaqueci mento de um grupo do Auto Scaling e otimizar a performan ce de políticas de escalabil idade que escalam continuam ente, habilitando o aquecimen to de instâncias padrão. Para obter mais informaçõ es, consulte Set the default instance warmup for an Auto Scaling group (Definir o aquecimento de instância padrão para um grupo do Auto Scaling) no Guia do usuário do Amazon EC2 Auto Scaling.	19 de abril de 2022

Alterações do guia	Um novo capítulo sobre integração com outros AWS serviços foi adicionado ao guia. Para obter mais informações, consulte AWS serviços integrados ao Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	29 de março de 2022
Atualizar permissões de perfil vinculado a serviços do IAM	A política `AutoScalingServiceRolePolicy` concede permissões de leitura adicionais ao Amazon EC2 Auto Scaling. Para obter mais informações, consulte Políticas gerenciadas pela AWS para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	28 de março de 2022
Os metadados da instância fornecem o estado de destino do ciclo de vida	É possível recuperar o estado de destino do ciclo de vida de uma instância do Auto Scaling nos metadados de instância . Para mais informações, consulte Retrieve the target lifecycle state through instance metadata (Recuperar o estado de destino do ciclo de vida por meio de metadados da instância) no Guia do usuário do Amazon EC2 Auto Scaling.	24 de março de 2022

Suporte à nova funcionalidade de pool ativo	Agora você pode hibernar instâncias em um pool ativo para interromper instâncias sem excluir o conteúdo da memória (RAM). Agora você também pode devolver instâncias ao pool ativo em redução de escala na horizontal, em vez de sempre terminar a capacidade da instância que você precisará posteriormente. Para obter mais informações, consulte Pools ativos para o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	24 de fevereiro de 2022
Alterações do guia	O console do Amazon EC2 Auto Scaling foi atualizado com outras opções para ajudar você a iniciar uma atualização de instância com ignorar correspondência habilitado e uma configuração desejada especificada. Para obter mais informações, consulte Iniciar ou cancelar uma atualização de instância (console) no Guia do usuário do Amazon EC2 Auto Scaling.	3 de fevereiro de 2022

Métricas personalizadas para políticas de escalabilidade preditiva	Agora, você pode escolher se deseja usar métricas personalizadas ao criar políticas de escalabilidade preditiva. Também é possível usar a métrica matemática para personalizar ainda mais as métricas incluídas na política. Para mais informações, consulte Advanced predictive scaling policy configurations using custom metrics (Configurações avançadas de política de escalabilidade preditiva usando métricas personalizadas).	24 de novembro de 2021
Nova estratégia de alocação sob demanda	Agora, é possível escolher se deseja iniciar instâncias sob demanda com base no preço (primeiro os tipos de instância com preços mais baixos) ao criar um grupo do Auto Scaling que usa uma política de instâncias mistas. Para mais informações, consulte Allocation strategies (Estratégias de alocação) no Guia do usuário do Amazon EC2 Auto Scaling.	27 de outubro de 2021

Seleção de tipo de instância baseada em atributos	O Amazon EC2 Auto Scaling adiciona suporte à seleção de tipo de instância baseada em atributos. Em vez de escolher manualmente os tipos de instância, você pode expressar seus requisitos de instância como um conjunto de atributos, como vCPU, memória e armazenamento. Para mais informações, consulte Creating an Auto Scaling group using attribute-based instance type selection (Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributo) no Guia do usuário do Amazon EC2 Auto Scaling.	27 de outubro de 2021
Suporte para filtragem de grupos por etiquetas	Agora você pode filtrar seus grupos do Auto Scaling usando filtros de etiquetas ao recuperar informaçõ es sobre grupos do Auto Scaling usando o comando `describe-auto-scal ing-groups` . Para mais informações, consulte Use tags to filter Auto Scaling groups (Usar etiquetas para filtrar grupos do Auto Scaling) no Guia do usuário do Amazon EC2 Auto Scaling.	14 de outubro de 2021

Alterações do guia	O console do Amazon EC2 Auto Scaling foi atualizado para ajudar você a criar políticas de rescisão personalizadas com. AWS Lambda A documentação do console foi adequadamente revisada. Para mais informações, consulte Using different termination policies (console) (Usar diferentes políticas de encerramento [console]).	14 de outubro de 2021
Suporte para cópia de configurações de execução para modelos de execução	Agora você pode copiar todas as configurações de lançamento em uma AWS região para novos modelos de lançamento do console do Amazon EC2 Auto Scaling.	9 de agosto de 2021

Expande a funcionalidade de atualização de instância	Agora você pode incluir atualizações, como uma nova versão de um modelo de execução, ao substitui r instâncias, adicionando a configuração desejada ao comando `start-ins tance-refresh` . Você também pode ignorar a substituição de instâncias que já têm a configuração desejada ativando a correspon dência de ignorar. Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na atualizaç ão de uma instância no Manual do usuário do Amazon EC2 Auto Scaling.	5 de agosto de 2021
Suporte para políticas de término personalizadas	Agora você pode criar políticas de rescisão personalizadas com AWS Lambda. Para obter mais informações, consulte Criação de uma política de término personalizada com o Lambda. A documentação para especificar políticas de término foi atualizada de maneira adequada.	29 de julho de 2021

Alterações do guia	O console do Amazon EC2 Auto Scaling foi atualizado e aprimorado com outros recursos para ajudar a criar ações programadas com um fuso horário especificado. A documentação para Escalabilidade programada foi revisada em conformidade.	3 de junho de 2021
volumes gp3 em configurações de execução	Agora você pode especificar volumes gp3 nos mapeamentos de dispositivos de bloco para configurações de execução.	2 de junho de 2021
Suporte para escalabilidade preditiva	Agora, você pode usar a escalabilidade preditiva para escalar proativamente grupos do Amazon EC2 Auto Scaling usando uma política de escalabilidade. Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling. Com essa atualização, a política AutoScalingServiceRolePolicygerenciada agora inclui permissão para chamar a ação `cloudwatch:GetMetricData` da API.	19 de maio de 2021

Alterações do guia	Agora você pode acessar modelos de exemplo para ganchos de ciclo de vida em. GitHub Para obter mais informações, consulte Ganchos do ciclo de vida do Auto Scaling do Amazon EC2 no Manual do usuário do Amazon EC2 Auto Scaling.	9 de abril de 2021
Suporte a pools ativos	Agora você pode equilibrar performance (minimizar inícios a de baixa atividade) e custo (interromper o provisionamento excessivo da capacidade da instância) para aplicações com longos primeiros tempos de inicialização adicionando pools ativos aos grupos do Auto Scaling. Para obter mais informações, consulte Pools ativos para o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	8 de abril de 2021

Suporte para pontos de verificação	Agora você pode adicionar pontos de verificação a uma atualização de instância para substituir instâncias em fases e executar verificações em suas instâncias em pontos específicos. Para obter mais informações, consulte Adição de pontos de verificação a uma atualização de instância no Manual do usuário do Amazon EC2 Auto Scaling.	18 de março de 2021
Alterações do guia	Documentação aprimorada para uso EventBridge com eventos e ganchos de ciclo de vida do Amazon EC2 Auto Scaling. Para obter mais informações, consulte Como usar o Amazon EC2 Auto Scaling EventBridge com um tutorial: Configurar um gancho de ciclo de vida que invoque uma função Lambda no Guia do usuário do Amazon EC2 Auto Scaling.	18 de março de 2021

Suporte para fusos horários locais	Agora você pode criar ações programadas recorrentes no fuso horário local adicionando a opção `--time-zone` ao comando put-scheduled-update-group-action. Se o seu fuso horário seguir o horário de verão, a ação recorrente ajustará automaticamente o horário de verão (DST). Para obter mais informações, consulte Escalabilidade programada no Manual do usuário do Amazon EC2 Auto Scaling.	9 de março de 2021
Expande a funcionalidade para políticas de instâncias mistas	Agora, você pode priorizar tipos de instância para sua capacidade spot quando usar uma política de instâncias mistas. O Amazon EC2 Auto Scaling tenta atender as prioridades com base no melhor esforço, mas primeiro otimiza a capacidade. Para obter mais informações, consulte Grupos de Auto Scaling com vários tipos de instância e opções de compra no Manual do usuário do Amazon EC2 Auto Scaling.	8 de março de 2021

Escalabilidade de atividades para grupos excluídos	Agora você pode visualizar atividades de escalabilidade para grupos do Auto Scaling excluídos adicionando a opção `--include-deleted-groups` ao comando describe-scaling-activities. Para obter mais informações, consulte Solução de problemas do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	23 de fevereiro de 2021
Melhorias no console	Agora você pode criar e anexar um Application Load Balancer ou Network Load Balancer ao console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte Criar e anexar um novo Application Load Balancer ou Network Load Balancer (console) no Guia do usuário do Amazon EC2 Auto Scaling.	24 de novembro de 2020
Várias interfaces de rede	Agora você pode configurar um modelo de execução para um grupo do Auto Scaling que especifique várias interfaces de rede. Para obter mais informações, consulte Interfaces de rede em uma VPC.	23 de novembro de 2020

Vários modelos de execução	Vários modelos de execução podem agora ser usados com grupos do Auto Scaling. Para obter mais informaçõ es, consulte Especificar um modelo de execução diferente para um tipo de instância no Manual do usuário do Amazon EC2 Auto Scaling.	19 de novembro de 2020
Balanceadores de carga de gateway	Guia atualizado para mostrar como anexar um balancead or de carga de gateway a um grupo do Auto Scaling para garantir que as instância s de dispositivo executada s pelo Amazon EC2 Auto Scaling sejam registradas automaticamente e cancelada s no balanceador de carga. Para obter mais informações, consulte Tipos de Elastic Load Balancing e Anexação de um balanceador de carga ao seu grupo do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	10 de novembro de 2020

Vida útil máxima da instância	Agora você pode reduzir a duração máxima da instância para um dia (86.400 segundos). Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na vida útil máxima da instância no Manual do usuário do Amazon EC2 Auto Scaling.	9 de novembro de 2020
Rebalanceamento de capacidade	Você pode configurar seu grupo do Auto Scaling para iniciar uma instância spot de substituição quando o Amazon EC2 emitir uma recomenda ção de rebalanceamento. Para obter mais informações, consulte Rebalanceamento de capacidade do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	4 de novembro de 2020
Instance Metadata Service Version 2	É possível usar o Instance Metadata Service Version 2, que é um método orientado a sessão para solicitação de metadados da instância ao usar as configurações de execução. Para obter mais informações, consulte Configurar as opções de metadados de instância no Manual do usuário do Amazon EC2 Auto Scaling.	28 de julho de 2020

Alterações do guia	Várias melhorias e novos procedimentos de console nas seções Controle de quais instâncias do Auto Scaling são terminadas durante a redução de escala na horizontal, Monitoramento das instâncias e grupos do Auto Scaling, Modelos de execução e Configurações de execução do Manual do usuário do Amazon EC2 Auto Scaling.	28 de julho de 2020
Atualização de instância	Inicie uma atualização de instância para atualizar todas as instâncias no seu grupo do Auto Scaling quando você fizer uma alteração de configuração. Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na atualização de uma instância no Manual do usuário do Amazon EC2 Auto Scaling.	16 de junho de 2020

Alterações do guia	Várias melhorias nas seções Substituição de instâncias do Auto Scaling com base no tempo de vida máximo da instância, Grupos do Auto Scaling com vários tipos de instância e opções de compra, Escalabilidade baseada no Amazon SQS e Marcação de instância e grupos do Auto Scaling do Manual do usuário do Amazon EC2 Auto Scaling.	6 de maio de 2020
Alterações do guia	Várias melhorias na documentação do IAM. Para obter mais informações, consulte Suporte a modelos de execução e Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	4 de março de 2020

Desabilitar políticas de escalabilidade	Agora você pode desabilitar e reabilitar as políticas de escalabilidade. Esse recurso permite desabilitar temporariamente uma política de escalabilidade enquanto preserva os detalhes de configuração para que você possa habilitar a política novamente mais tarde. Para obter mais informações, consulte Desativação de uma política de escalabilidade para um grupo do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	18 de fevereiro de 2020
Adicionar funcionalidade de notificação	O Amazon EC2 Auto Scaling agora envia eventos para AWS Health Dashboard você quando seus grupos do Auto Scaling não conseguem escalar devido à falta de um grupo de segurança ou modelo de lançamento. Para obter mais informaçõ es, consulte Notificações do Health Dashboard para o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	12 de fevereiro de 2020

Alterações do guia	Várias melhorias e correções nas seções Como o Amazon EC2 Auto Scaling funciona com o IAM, Exemplo de políticas baseadas em identidade do Amazon EC2 Auto Scaling, Política de chave da CMK necessári a para uso com volumes criptografados e Monitoram ento das suas instâncias e grupos do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	10 de fevereiro de 2020
Alterações do guia	Melhoria na documentação para grupos do Auto Scaling que usam peso de instância s. Saiba como usar políticas de escalabilidade ao usar “unidades de capacidade” para medir a capacidade desejada. Para obter mais informações, consulte Como funcionam as políticas de escalabilidade e Tipos de ajuste da escalabilidade no Manual do usuário do Amazon EC2 Auto Scaling.	6 de fevereiro de 2020

Novo capítulo “Segurança”	Um novo capítulo sobre Segurança no Guia do usuário do Amazon EC2 Auto Scaling ajuda você a entender como aplicar o modelo de responsabilidade compartilhada ao usar o Amazon EC2 Auto Scaling. Como parte dessa atualização, o capítulo do Manual do usuário "Controle de acesso aos recursos do Amazon EC2 Auto Scaling" foi substituído por uma seção nova e mais útil, Gerenciamento de identidades e acesso para o Amazon EC2 Auto Scaling.	4 de fevereiro de 2020
Recomendações para tipos de instância	AWS Compute Optimizer fornece recomendações de instâncias do Amazon EC2 para ajudar você a melhorar o desempenho, economizar dinheiro ou ambos. Para obter mais informações, consulte Obtenção de recomendações de um tipo de instância no Manual do usuário do Amazon EC2 Auto Scaling.	3 de dezembro de 2019

Hosts dedicados e grupos de recursos de host	Guia atualizado para mostrar como criar um modelo de execução que especifica um grupo de recursos de host. Isso permite criar um grupo do Auto Scaling com um modelo de execução que especific a uma AMI de BYOL a ser usada em hosts dedicados . Para obter mais informaçõ es, consulte Criação de um modelo de execução para um grupo do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	3 de dezembro de 2019
Suporte a endpoints da Amazon VPC	Agora você pode estabelec er uma conexão privada entre sua VPC e o Amazon EC2 Auto Scaling. Para obter mais informações, consulte Amazon EC2 Auto Scaling e endpoints da VPC da interface no Manual do usuário do Amazon EC2 Auto Scaling.	22 de novembro de 2019

Vida útil máxima da instância	Agora você pode substituir instâncias de forma automática especificando o período máximo que uma instância pode estar em serviço. Se alguma instância estiver se aproximando desse limite, o Amazon EC2 Auto Scaling gradualmente as substituirá. Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na vida útil máxima da instância no Manual do usuário do Amazon EC2 Auto Scaling.	19 de novembro de 2019
Ponderação de instâncias	Para grupos do Auto Scaling com vários tipos de instância, agora você pode especificar opcionalmente o número de unidades de capacidade com que cada tipo de instância contribui para a capacidade do grupo. Para obter mais informações, consulte Ponderação de instâncias do Auto Scaling do Amazon EC2 no Manual do usuário do Amazon EC2 Auto Scaling.	19 de novembro de 2019

Número mínimo de tipos de instância	Você não precisa mais especificar tipos de instância adicionais para grupos de instâncias spot, sob demanda e reservadas. Para todos os grupos de Auto Scaling, o mínimo agora é um tipo de instância. Para obter mais informações, consulte Grupos de Auto Scaling com vários tipos de instância e opções de compra no Manual do usuário do Amazon EC2 Auto Scaling.	16 de setembro de 2019
Suporte para a nova estratégia de alocação spot	O Amazon EC2 Auto Scaling agora oferece suporte a uma nova estratégia de alocação spot "otimizada para capacidade" que atende à sua solicitação usando grupos de instâncias spot escolhidos de forma ideal com base na capacidade spot disponível. Para obter mais informações, consulte Grupos de Auto Scaling com vários tipos de instância e opções de compra no Manual do usuário do Amazon EC2 Auto Scaling.	12 de agosto de 2019

Alterações do guia	Documentação do Amazon EC2 Auto Scaling melhorada nos tópicos Perfis vinculados ao serviço e Política de chaves de CMK necessária para uso com volumes criptografados.	1 de agosto de 2019
Suporte para aprimoramento de marcação	Agora, o Amazon EC2 Auto Scaling adiciona tags às instâncias do Amazon EC2 como parte da mesma chamada de API que inicia as instâncias. Para obter mais informações, consulte Marcação de grupos e instâncias do Auto Scaling.	26 de julho de 2019
Alterações do guia	Melhoria na documentação do Amazon EC2 Auto Scaling no tópico Suspensão e retomada de processos de escalabilidade. Atualização nos Exemplos de políticas gerenciadas pelo cliente para incluir um exemplo de política que permite que os usuários transmitam apenas perfis vinculados ao serviço com sufixo personalizado específicas para o Amazon EC2 Auto Scaling.	13 de junho de 2019

Suporte para novos recursos do Amazon EBS	Adicionado suporte para novos recursos do Amazon EBS no tópico de modelo de execução. Altere o estado de criptografia de um volume do EBS ao restaurar um snapshot. Para obter mais informações, consulte Criação de um modelo de execução para um grupo do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	13 de maio de 2019
Alterações do guia	Melhoria na documenta ção do Amazon EC2 Auto Scaling nas seguintes seções: Controle de quais instâncias do Auto Scaling são terminada s durante uma redução de escala na horizontal, Grupos do Auto Scaling, Grupos do Auto Scaling com vários tipos de instâncias e opções de compra e Escalabilidade dinâmica para o Amazon EC2 Auto Scaling.	12 de março de 2019

Suporte para a combinação de tipos de instâncias e opções de compra	Provisione e escale instâncias automaticamente nas opções de compra (spot, sob demanda e instâncias reservadas) e tipos de instância em um único grupo do Auto Scaling. Para obter mais informações, consulte Grupos de Auto Scaling com vários tipos de instância e opções de compra no Manual do usuário do Amazon EC2 Auto Scaling.	13 de novembro de 2018
Tópico atualizado para escalabilidade baseada no Amazon SQS	Guia atualizado para explicar como você pode usar métricas personalizadas para escalar um grupo do Auto Scaling em resposta à mudança na demanda de uma fila do Amazon SQS. Para obter mais informações, consulte Escalabilidade baseada no Amazon SQS no Manual do usuário do Amazon EC2 Auto Scaling.	26 de julho de 2018

A tabela a seguir descreve alterações importantes feitas na documentação do Amazon EC2 Auto Scaling antes de julho de 2018.

Recurso	Description	Data de lançamento
Suporte para as políticas de escalabilidade	Configure a escalabilidade dinâmica da sua aplicação em apenas algumas etapas. Para obter mais informações,	12 de julho de 2017

Recurso	Description	Data de lançamento
de rastreamento de destino	consulte Políticas de escalabilidade com monitoramento do objetivo do Amazon EC2 Auto Scaling.	
Suporte para permissões em nível de recurso	Criar políticas do IAM para controlar acesso em nível de recurso. Para obter mais informações, consulte Controle do acesso aos seus recursos do Amazon EC2 Auto Scaling.	15 de maio de 2017
Monitoramento de melhorias	As métricas de grupo do Auto Scaling não precisam mais que você habilite monitoramento detalhado. Agora você pode habilitar a coleta de métricas do grupo e visualizar gráficos de métricas na guia Monitoramento no console. Para obter mais informações, consulte Monitoramento de seus grupos e instâncias do Auto Scaling usando a Amazon. CloudWatch	18 de agosto de 2016
Suporte para Application Load Balancers	Anexar um ou mais grupos de destino a um grupo novo ou existente do Auto Scaling. Para obter mais informaçõ es, consulte Anexação de um balanceador de carga ao seu grupo do Auto Scaling.	11 de agosto de 2016
Eventos para ganchos do ciclo de vida	O Amazon EC2 Auto Scaling envia eventos EventBrid ge para quando chama ganchos de ciclo de vida. Para obter mais informações, consulte Como obter EventBrid ge quando seu grupo de Auto Scaling é dimensionado.	24 de fevereiro de 2016
Proteção de instância	Impedir que o Amazon EC2 Auto Scaling selecione instâncias específicas para término ao reduzir a escala. Para obter mais informações, consulte Proteção de instância.	07 de dezembro de 2015
Políticas de escalabilidade em etapas	Criar uma política de escalabilidade que permita escalonar com base no tamanho da violação do alarme. Para obter mais informações, consulte Tipos de política de escalabilidade.	06 de julho de 2015

Recurso	Description	Data de lançamento
Atualizar balanceador de carga	Anexar ou desvincular um balanceador de carga de um grupo do Auto Scaling existente. Para obter mais informações, consulte Anexação de um balanceador de carga ao seu grupo do Auto Scaling.	11 de junho de 2015
Support for ClassicLink	Vincular instâncias EC2-Classic em seu grupo do Auto Scaling a uma VPC permitindo comunicação entre essas instâncias EC2-Classic vinculadas e instâncias na VPC usando endereços IP privados. Para obter mais informações, consulte Vinculação de instâncias do EC2-Classic a uma VPC.	19 de janeiro de 2015
Ganchos do ciclo de vida	Manter suas instâncias recém-ativadas ou encerradas em um estado pendente enquanto você realiza ações nelas. Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.	30 de julho de 2014
Desvincular instâncias	Desvincular instâncias de um grupo do Auto Scaling. Para obter mais informações, consulte Desvincular instâncias do EC2 do seu grupo do Auto Scaling.	30 de julho de 2014
Colocar instâncias em um estado de Standby	Colocar instâncias que estão em um estado de `InService` em um estado de `Standby`. Para obter mais informações, consulte Remoção temporária de instâncias do seu grupo do Auto Scaling.	30 de julho de 2014
Gerenciamento de tags	Gerencie seus grupos do Auto Scaling usando o Console de gerenciamento da AWS. Para obter mais informações, consulte Marcação de grupos e instâncias do Auto Scaling.	01 de maio de 2014
Suporte para instâncias dedicadas	Ativar instâncias dedicadas especificando um atributo de locação de localização ao criar uma configuração de ativação. Para obter mais informações, consulte Locação de posicionamento de instância.	23 de abril de 2014

Recurso	Description	Data de lançamento
Criar um grupo ou configuração de ativação a partir de uma instância do EC2	Criar um grupo do Auto Scaling ou uma configuração de execução usando uma instância do EC2. Para obter informações sobre como criar uma configuração de execução usando uma instância do EC2, consulte Criação de uma configuração de execução usando uma instância do EC2. Para obter informações sobre como criar um grupo do Auto Scaling usando uma instância do EC2, consulteCriação de um grupo do Auto Scaling usando uma instância do EC2.	02 de janeiro de 2014
Anexar instâncias	Habilite a escalabilidade automática para uma instância do EC2 anexando a instância a um grupo do Auto Scaling existente. Para obter mais informações, consulte Anexar instâncias do EC2 ao seu grupo do Auto Scaling.	02 de janeiro de 2014
Visualizar limites da conta	Visualize os limites dos recursos do Auto Scaling para a sua conta. Para obter mais informações, consulte Cotas para recursos e grupos do Auto Scaling.	02 de janeiro de 2014
Suporte para console do Amazon EC2 Auto Scaling	Acesse o Amazon EC2 Auto Scaling usando o. Console de gerenciamento da AWS Para obter mais informações, consulte Conceitos básicos do Amazon EC2 Auto Scaling.	10 de dezembro de 2013
Atribuir um endereço IP público	Atribuir um endereço IP público a uma instância iniciada em uma VPC. Para obter mais informações, consulte Execução de instâncias do Auto Scaling em uma VPC.	19 de setembro de 2013
Política de encerramento de instância	Especificar uma política de término de instância para o Amazon EC2 Auto Scaling usar ao terminar instâncias do EC2. Para obter mais informações, consulte Controle de quais instâncias do Auto Scaling são terminadas durante uma redução de escala na horizontal.	17 de setembro de 2012

Recurso	Description	Data de lançamento
Suporte para perfis do IAM	Iniciar instâncias do EC2 com um perfil de instância do IAM. Você pode usar esse recurso para atribuir perfis do IAM a suas instâncias, permitindo que suas aplicações acessem outros Amazon Web Services com segurança. Para obter mais informações, consulte Ativar instâncias do Auto Scaling com um perfil do IAM.	11 de junho de 2012
Suporte a instâncias spot	Execução de instâncias spot com uma configuração de execução. Para mais informações, consulte Requesting Spot Instances for fault-tolerant and flexible applications (Solicitar instâncias spot para aplicações flexíveis e tolerantes a falhas).	7 de junho de 2012
Marcar grupos e instâncias	Marcar grupos do Auto Scaling e especificar se a tag também se aplica a instâncias do EC2 iniciadas depois que a tag foi criada. Para obter mais informações, consulte Marcação de grupos e instâncias do Auto Scaling.	26 de janeiro de 2012

Recurso	Description	Data de lançamento
Suporte para o Amazon SNS	Use o Amazon SNS para receber notificações sempre que o Amazon EC2 Auto Scaling iniciar ou terminar instâncias do EC2. Para obter mais informações, consulte Obtenção de notificações do SNS quando o grupo do Auto Scaling é escalado. O Amazon EC2 Auto Scaling também adicionou os seguintes novos recursos: • A capacidade de configurar ações de escalabilidade recorrentes usando a sintaxe cron. Para obter mais informações, consulte a operação da API `PutScheduledUpdateGroupAction`. • Uma nova configuração que permite escalar sem adicionar a instância executada ao balanceador de carga (LoadBalancer). Para obter mais informações, consulte o tipo de dados da API `ProcessType`. • O sinalizador `ForceDelete` na operação `DeleteAutoScalingGroup` que informa ao Amazon EC2 Auto Scaling para excluir o grupo do Auto Scaling com as instâncias associadas a ele sem esperar que as instâncias sejam terminadas primeiro. Para obter mais informações, consulte a operação da API `DeleteAutoScalingGroup`.	20 de julho de 2011
Ações de escalabilidade programadas	Suporte adicional para ações de escalabilidade programadas. Para obter mais informações, consulte Escalabilidade programada do Amazon EC2 Auto Scaling.	2 de dezembro de 2010
Suporte para a Amazon VPC	Adicionado suporte para a Amazon VPC. Para obter mais informações, consulte Execução de instâncias do Auto Scaling em uma VPC.	2 de dezembro de 2010

Recurso	Description	Data de lançamento
O suporte a clusters HPC	Suporte adicionado para clusters de computação de alta performance (HPC).	2 de dezembro de 2010
Suporte a verificações de integridade	Suporte adicionado para o uso de verificações de integridade do Elastic Load Balancing com instâncias do EC2 gerenciadas pelo Amazon EC2 Auto Scaling. Para obter mais informações, consulte Verificações de saúde de instâncias em um grupo de Auto Scaling.	2 de dezembro de 2010
Support para CloudWatch alarmes	O mecanismo de acionamento antigo foi removido e o Amazon EC2 Auto Scaling foi redesenhado para usar o recurso de alarme. CloudWatch Para obter mais informações, consulte Escalabilidade dinâmica do Amazon EC2 Auto Scaling.	2 de dezembro de 2010
Suspender e retomar a escalabilidade	Suporte adicional para suspender e retomar processos de escalabilidade.	2 de dezembro de 2010
Suporte ao IAM	Adicionado suporte ao IAM. Para obter mais informaçõ es, consulte Controle do acesso aos seus recursos do Amazon EC2 Auto Scaling.	2 de dezembro de 2010

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.